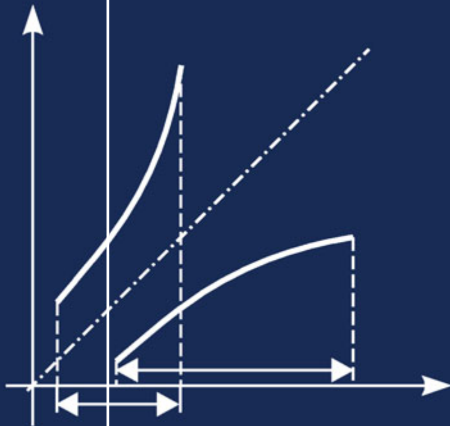


I. N. Bronshtein   K. A. Semendyayev
G. Musiol   H. Mühlig

HANDBOOK OF MATHEMATICS

Sixth Edition

Springer

Handbook of Mathematics

I.N. Bronshtein · K.A. Semendyayev
Gerhard Musiol · Heiner Mühlig

Handbook of Mathematics

Sixth Edition

With 799 Figures and 132 Tables

Springer

I.N. Bronshtein (Deceased)

K.A. Semendyayev (Deceased)

Gerhard Musiol
Dresden, Sachsen
Germany

Heiner Mühlig
Dresden, Sachsen
Germany

ISBN 978-3-662-46220-1 ISBN 978-3-662-46221-8 (eBook)
DOI 10.1007/978-3-662-46221-8

Library of Congress Control Number: 2015933616

Springer Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer-Verlag GmbH Berlin Heidelberg is part of Springer Science+Business Media
(www.springer.com)

Preface to the Sixth English Edition

This sixth English edition is based on the fifth English edition (2007) and corresponds to the improved seventh (2008), eighth (2012) and ninth (2013) German edition. It contains all the chapters of the mentioned editions, but in a renewed, revised and extended form (see also the preface to the fifth English edition).

Special new parts to be mentioned here are such supplementary sections as Geometric and Coordinate Transformations and Plain Projections in Chapter on Geometry, as Quaternions and Applications in Chapter on Linear Algebra, as Lie Groups and Lie Algebras in Chapter on Algebra and Discrete Mathematics and as Matlab in Chapter on Numerical Analysis. The Chapter on Computer Algebra Systems is restricted to Mathematica only as a representative example for such systems.

Extended and revised paragraphs are given in Chapter on Geometry about Cardan and Euler angles and about Coordinate transformations, in Chapter on Integral Calculus about Applications of Definite Integrals, in Chapter on Optimization about Evaluation Strategies, in Chapter on Tables about Natural Constants and Physical Units (System SI). The Index has been completed to an extent as in the previous German editions.

We would like to cordially thank all readers and professional colleagues who helped us with their valuable statements, remarks and suggestions on the German editions of the book during the revision process. Special thanks go to Mrs. Professor Dr. Gabriella Szép (Budapest), who made this English edition possible by valuable contributions and the basic translation into the English. Furthermore our thanks go to all co-authors for the critical treatment of their chapters.

Dresden, December 2014

Prof. Dr. Gerhard Musiol   Prof. Dr. Heiner Mühlig

Preface to the Fifth English Edition

This fifth edition is based on the fourth English edition (2003) and corresponds to the improved sixth German edition (2005). It contains all the chapters of the both mentioned editions, but in a renewed revised and extended form.

So in the work at hand, the classical areas of Engineering Mathematics required for current practice are presented, such as “Arithmetic”, “Functions”, “Geometry”, “Linear Algebra”, “Algebra and Discrete Mathematics”, (including “Logic”, “Set Theory”, “Classical Algebraic Structures”, “Finite Fields”, “Elementary Number Theory”, ”Cryptology”, “Universal Algebra”, “Boolean Algebra and Switch Algebra”, “Algorithms of Graph Theory”, “Fuzzy Logic”), “Differentiation”, “Integral Calculus”, “Differential Equations”, “Calculus of Variations”, “Linear Integral Equations”, “Functional Analysis”, “Vector Analysis and Vector Fields”, “Function Theory”, “Integral Transformations”, “Probability Theory and Mathematical Statistics”.

Fields of mathematics that have gained importance with regards to the increasing mathematical modeling and penetration of technical and scientific processes also receive special attention. Included amongst these chapters are “Stochastic Processes and Stochastic Chains” as well as “Calculus of Errors”, “Dynamical Systems and Chaos”, “Optimization”, “Numerical Analysis”, “Using the Computer” and “Computer Algebra Systems”.

The Chapter 21 containing a large number of useful tables for practical work has been completed by

adding tables with the physical units of the International System of Units (SI).

Dresden, February 2007

Prof. Dr. Gerhard Musiol — Prof. Dr. Heiner Mühlig

From the Preface to the Fourth English Edition

The "Handbook of Mathematics" by the mathematician, I. N. Bronshtein and the engineer, K. A. Semendyayev was designed for engineers and students of technical universities. It appeared for the first time in Russian and was widely distributed both as a reference book and as a text book for colleges and universities. It was later translated into German and the many editions have made it a permanent fixture in German-speaking countries, where generations of engineers, natural scientists and others in technical training or already working with applications of mathematics have used it.

On behalf of the publishing house Harri Deutsch, a revision and a substantially enlarged edition was prepared in 1992 by Gerhard Musiol and Heiner Mühlig, with the goal of giving "Bronshtein" the modern practical coverage requested by numerous students, university teachers and practitioners. The original style successfully used by the authors has been maintained. It can be characterized as "short, easily understandable, comfortable to use, but featuring mathematical accuracy (at a level of detail consistent with the needs of engineers)"*. Since 2000, the revised and extended fifth German edition of the revision has been on the market. Acknowledging the success that "Bronstein" has experienced in the German-speaking countries, Springer Verlag Heidelberg/Germany is publishing a fourth English edition, which corresponds to the improved and extended fifth German edition.

The book is enhanced with over a thousand complementary illustrations and many tables. Special functions, series expansions, indefinite, definite and elliptic integrals as well as integral transformations and statistical distributions are supplied in an extensive appendix of tables.

In order to make the reference book more effective, clarity and fast access through a clear structure were the goals, especially through visual clues as well as by a detailed technical index and colored tabs.

An extended bibliography also directs users to further resources.

Special thanks go to Mrs. Professor Dr. Gabriella Szép (Budapest), who made this English debut version possible.

Dresden, June 2003

Prof. Dr. Gerhard Musiol — Prof. Dr. Heiner Mühlig

Co-Authors

Some chapters or sections are the result of cooperation with co-authors.

Chapter resp. section	Co-author
Spherical Trigonometry (3.4.1 – 3.4.3.3)	Dr. H. Nickel †, Dresden
Spherical Curves (3.4.3.4)	Prof. L. Marsolek, Berlin
Geometric Transformations, Coordinate Transformations, Planar Projections (3.5.4, 3.5.5)	Dr. I. Steinert, Düsseldorf
Quaternions and Applications (4.4)	PD Dr. S. Bernstein, Freiberg (Sachsen)

*See Preface to the First Russian Edition

Logic (5.1), Set Theory (5.2), Classical Algebraic Structures (5.3), Applications of Groups (beyond) 5.3.4, 5.3.5.4 – 5.3.5.6), Rings and Fields (5.3.7), Vector Spaces (5.3.8), Boolean Algebra and Switch Algebra (5.7), Universal Algebra (5.6)	Dr. J. BRUNNER, Dresden
Group Representation (5.3.4), Applications of Groups (5.3.5.4 – 5.3.5.6)	Prof. Dr. R. REIF, Dresden
LIE–Groups and LIE–Algebras (5.3.6)	PD Dr. S. BERNSTEIN, Freiberg (Sachsen)
Elementary Number Theory (5.4), Cryptology , (5.5) Graphs (5.8)	Prof. Dr. U. BAUMANN, Dresden
Fuzzy–Logic (5.9)	Prof. Dr. A. GRAUEL, Soest
Important Formulas for the Spherical Bessel Functions (9.1.2.6, sub-point 2.5)	Prof. Dr. P. Ziesche, Dresden
Statistical Interpretation of the Wave Function (9.2.4.4)	Prof. Dr. R. REIF, Dresden
Non-linear partial Differential Equations: Solitons, Periodic Patterns and Chaos (9.2.5)	Prof. Dr. P. ZIESCHE, Dresden
Dissipative Solitons, Light and Dark Solitons (9.2.5.3, in point 2)	Dr. J. Brand, Dresden
Linear Integral Equations (11)	Dr. I. STEINERT, Düsseldorf
Functional analysis (12)	Prof. Dr. M. WEBER, Dresden
Elliptic Functions (14.6)	Dr. N. M. FLEISCHER †, Moskau
Dynamical Systems and Chaos (17)	Prof. Dr. V. REITMANN, St. Petersburg
Optimization (18)	Dr. I. STEINERT, Düsseldorf
Using the Computer: (19.8.1, 19.8.2), Interactive System: Mathematica (19.8.4.2), Maple (19.8.4.3), Computeralgebra Systems – Example Mathematica (20)	Prof. Dr. G. FLACH, Dresden
Interactive System: Matlab (19.8.4.1)	PD Dr. B. Mulansky, Clausthal
Computeralgebra Systems – Example Mathematica (20): Revision of the chapter in accordance with version 10 of Mathematica	Dr. J. Tóth, Budapest

Additional Chapters with Co-Authors in the CD–ROM to the Books of the German Editions 7,8 and 9.

Lie–Groups and Lie–Algebras (5.3.5), (5.3.6)	Prof. Dr. R. Reif, Dresden
Non-linear Partial Differential Equations: Inverse Scattering Theory (methods in analogy to the Fourier method)(9.2.6)	Dr. B. Rumpf,
Mathematical Basis of Quantum Mechanics (21)	Prof. Dr. A. Buchleitner, PD Dr. M. Tiersch, Dr. Th. Wellens, Freiburg
Quantencomputer (22)	Prof. Dr. A. Buchleitner, PD Dr. M. Tiersch, Dr. Th. Wellens, Freiburg

Contents

List of Tables

1 Arithmetics

1.1 Elementary Rules for Calculations

1.1.1 Numbers

1.1.1.1 Natural, Integer, and Rational Numbers

1. Definitions and Notation

The positive and negative integers, fractions, and zero together are called the *rational numbers*. In relation to these the following notations are used (see 5.2.1, **1.**, p. 327):

- Set of natural numbers: $\mathbb{N}=\{0, 1, 2, 3, \ldots\}$,
- Set of integers: $\mathbb{Z}= \{\ldots, -2, -1, 0, 1, 2, \ldots\}$,
- Set of rational numbers: $\mathbb{Q}= \{x|x = \frac{p}{q} \text{ with } p \in \mathbb{Z},\ q \in \mathbb{Z} \text{ and } q \neq 0\}$.

The notion of natural numbers arose from enumeration and ordering. The natural numbers are also called the *non-negative integers*.

2. Properties of the Set of Rational Numbers

- The set of rational numbers is infinite.
- The set is *ordered*, i.e., for any two different given numbers a and b one can tell which is the smaller one.
- The set is *dense everywhere*, i.e., between any two different rational numbers a and b $(a < b)$ there is at least one rational number c $(a < c < b)$. Consequently, there is an infinite number of other rational numbers between any two different rational numbers.

3. Arithmetical Operations

The arithmetical operations (addition, subtraction, multiplication and division) can be performed with any two rational numbers, and the result is a rational number. The only exception is *division by zero*, which is not possible: The operation written in the form $a : 0$ is meaningless because it does not have any result: If $a \neq 0$, then there is no rational number b such that $b \cdot 0 = a$ could be fulfilled, and if $a = 0$ then b can be any of the rational numbers. The frequently occurring formula $a : 0 = \infty$ (infinity) does not mean that the division is possible; it is only the notation for the statement: If the denominator approaches zero and, e.g., the numerator does not, then the absolute value (magnitude) of the quotient exceeds any finite limit.

4. Decimal Fractions, Continued Fractions

Every rational number a can be represented as a terminating or periodically infinite decimal fraction or as a finite continued fraction (see 1.1.1.4, p. 3).

5. Geometric Representation

Fixing an *origin* the *zero point* 0, a positive direction the *orientation*, and the unit of length l the *measuring rule*, (see also 2.17.1, p. 115 and **(Fig. 1.1)**), then every rational number a corresponds to a certain point on this line. This point has the coordinate a, and it is a so-called *rational point*. The line is called the *numerical axis*. Because the set of rational numbers is dense everywhere, between two rational points there are infinitely many further rational points.



Figure 1.1

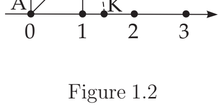


Figure 1.2

I.N. Bronshtein et al., *Handbook of Mathematics*,
DOI 10.1007/978-3-662-46221-8_1

1.1.1.2 Irrational and Transcendental Numbers

The set of rational numbers is not satisfactory for calculus. Even though it is dense everywhere, it does not cover the whole numerical axis. If for example the diagonal AB of the unit square rotates around A so that B goes into the point K, then K does not have any rational coordinate **(Fig. 1.2)**.
The introduction of *irrational numbers* allows to assign a number to every point of the numerical axis. In textbooks there are given exact definitions for irrational numbers, e.g., by nests of intervals. For this survey it is enough to note that the irrational numbers take all the non-rational points of the numerical axis and every irrational number corresponds to a point of the axis, and that every irrational number can be represented as a non-periodic infinite decimal fraction.
First of all, the non-integer real roots of the algebraic equation

$$x^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0 = 0 \quad (n > 1,\ \text{integer; integer coefficients}), \tag{1.1a}$$

belong to the irrational numbers. These roots are called *algebraic irrationals*.

■ **A:** The simplest examples of algebraic irrationals are the real roots of $x^n - a = 0$ $(a > 0)$, as numbers of the form $\sqrt[n]{a}$, if they are not rational.

■ **B:** $\sqrt[2]{2} = 1.414\ldots$, $\sqrt[3]{10} = 2.154\ldots$ are algebraic irrationals.

The irrational numbers which are not algebraic irrationals are called *transcendental*.

■ **A:** $\pi = 3.141592\ldots$ $e = 2.718281\ldots$ are transcendental numbers.

■ **B:** The decimal logarithm of the integers, except the numbers of the form 10^n, are transcendental.

The non-integer roots of the quadratic equation

$$x^2 + a_1x + a_0 = 0 \quad (a_1, a_0\ \text{integers}) \tag{1.1b}$$

are called *quadratic irrationals*. They have the form $(a + b\sqrt{D})/c$ $(a, b, c$ integers, $c \neq 0$; $D > 0$, square-free number).

■ The division of a line segment a in the ratio of the golden section $x/a = (a - x)/x$ (see 3.5.2.3, **3.**, p. 194) leads to the quadratic equation $x^2 + x - 1 = 0$, if $a = 1$. The solution $x = (\sqrt{5} - 1)/2$ is a quadratic irrational. It contains the irrational number $\sqrt{5}$.

1.1.1.3 Real Numbers

Rational and irrational numbers together form the *set of real numbers*, which is denoted by $\mathbb{R}$.

1. Most Important Properties

The set of real numbers has the following important properties (see also 1.1.1.1, **2.**, p. 1). It is:
- *Infinite.*
- *Ordered.*
- *Dense everywhere.*
- *Closed*, i.e., every point of the numerical axis corresponds to a real number. This statement does not hold for the rational numbers.

2. Arithmetical Operations

Arithmetical operations can be performed with any two real numbers and the result is a real number, too. The only exception is division by zero (see 1.1.1.1, **3.**, p. 1). Raising to a power and also its inverse operation can be performed among real numbers; so it is possible to take an arbitrary root of any positive number; every positive real number has a logarithm for an arbitrary positive basis, except that 1 cannot be a basis.

A further generalization of the notion of numbers leads us to the concept of *complex numbers* (see 1.5, p. 34).

3. Interval of Numbers

A connected set of real numbers with endpoints a and b is called an *interval of numbers with endpoints a and b*, where $a < b$ and a is allowed to be $-\infty$ and b is allowed to be $+\infty$. If the endpoint itself does not belong to the interval, then this end of the interval is *open*, in the opposite case it is *closed*.

An interval is given by its endpoints a and b, putting them in braces: A bracket for a closed end of the interval and a parenthesis for an open one. It is to be distinguished between *open intervals* (a, b), *half-open (half-closed) intervals* $[a, b)$ or $(a, b]$ and *closed intervals* $[a, b]$, according to whether none of the endpoints, one of the endpoints or both endpoints belong to it, respectively. Frequently the notation $]a, b[$ instead of (a, b) for open intervals, and analogously $[a, b[$ instead of $[a, b)$ is used. In the case of graphical representations, in this book the open end of the interval is denoted by a round arrow head, the closed one by a filled point.

1.1.1.4 Continued Fractions

Continued fractions are nested fractions, by which rational and irrational numbers can be represented and approximated even better than by decimal representation (see 19.8.1.1, p. 1002 and ■ **A** and ■ **B** on p. 4).

1. Rational Numbers

The continued fraction of a rational number is finite. Positive rational numbers which are greater than 1 have the form (1.2). For abbreviation the symbol $\frac{p}{q} = [a_0; a_1, a_2, \ldots, a_n]$ is used with $a_k \geq 1 \ (k = 1, 2, \ldots, n)$.

$$\frac{p}{q} = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{\ddots + \cfrac{1}{a_{n-1} + \cfrac{1}{a_n}}}}} \,. \tag{1.2}$$

The numbers a_k are calculated with the help of the *Euclidean algorithm*:

$$\frac{p}{q} = a_0 + \frac{r_1}{q} \quad \left(0 < \frac{r_1}{q} < 1\right), \tag{1.3a}$$

$$\frac{q}{r_1} = a_1 + \frac{r_2}{r_1} \quad \left(0 < \frac{r_2}{r_1} < 1\right), \tag{1.3b}$$

$$\frac{r_1}{r_2} = a_2 + \frac{r_3}{r_2} \quad \left(0 < \frac{r_3}{r_2} < 1\right), \tag{1.3c}$$

$$\vdots \qquad \vdots \qquad \vdots$$

$$\frac{r_{n-2}}{r_{n-1}} = a_{n-1} + \frac{r_n}{r_{n-1}} \quad \left(0 < \frac{r_n}{r_{n-1}} < 1\right), \tag{1.3d}$$

$$\frac{r_{n-1}}{r_n} = a_n \quad (r_{n+1} = 0)\,. \tag{1.3e}$$

■ $$\frac{61}{27} = 2 + \frac{7}{27} = 2 + \cfrac{1}{3 + \cfrac{6}{7}} = 2 + \cfrac{1}{3 + \cfrac{1}{1 + \cfrac{1}{6}}} = [2; 3, 1, 6]\,.$$

2. Irrational Numbers

Continued fractions of irrational numbers do not break off. They are called infinite continued fractions with $[a_0; a_1, a_2, \ldots]$.

If some numbers a_k are repeated in an infinite continued fraction, then this fraction is called a *periodic continued fraction* or *recurring chain fraction*. Every periodic continued fraction represents a quadratic irrationality, and conversely, every quadratic irrationality has a representation in the form of a periodic continued fraction.

■ The number $\sqrt{2} = 1.4142135\ldots$ is a quadratic irrationality and it has the periodic continued fraction representation $\sqrt{2} = [1; 2, 2, 2, \ldots]$.

3. Aproximation of Real Numbers

If $\alpha = [a_0; a_1, a_2, \ldots]$ is an arbitrary real number, then every finite continued fraction

$$\alpha_k = [a_0; a_1, a_2, \ldots, a_k] = \frac{p}{q} \tag{1.4}$$

represents an approximation of α. The continued fraction α_k is called the k-th approximant of α. It can be calculated by the recursive formula

$$\alpha_k = \frac{p_k}{q_k} = \frac{a_k p_{k-1} + p_{k-2}}{a_k q_{k-1} + q_{k-2}} \quad (k \geq 1; \;\; p_{-1} = 1, p_0 = a_0; q_{-1} = 0, q_0 = 1). \tag{1.5}$$

According to the *Liouville approximation theorem*, the following estimat holds:

$$|\alpha - \alpha_k| = |\alpha - \frac{p_k}{q_k}| < \frac{1}{q_k^2}\,. \tag{1.6}$$

Furthermore, it can be shown that the approximants approach the real number α with increasing accuracy alternatively from above and from below. The approximants converge to α especially fast if the numbers a_i $(i = 1, 2, \ldots, k)$ in (1.4) have large values. Consequently, the convergence is worst for the numbers $[1; 1, 1, \ldots]$.

■ **A:** From the decimal presentation of π the continued fraction representation $\pi = [3; 7, 15, 1, 292, \ldots]$ follows with the help of (1.3a)–(1.3e). The corresponding approximants (1.5) with the estimate according to (1.6) are: $\alpha_1 = \frac{22}{7}$ with $|\pi - \alpha_1| < \frac{1}{7^2} \approx 2 \cdot 10^{-2}$, $\alpha_2 = \frac{333}{106}$ with $|\pi - \alpha_2| < \frac{1}{106^2} \approx 9 \cdot 10^{-5}$, $\alpha_3 = \frac{355}{113}$ with $|\pi - \alpha_3| < \frac{1}{113^2} \approx 8 \cdot 10^{-5}$. The actual errors are much smaller. They are less than $1.3 \cdot 10^{-3}$ for α_1, $8.4 \cdot 10^{-5}$ for α_2 and $2.7 \cdot 10^{-7}$ for α_3. The approximants α_1, α_2 and α_3 represent better approximations for π than the decimal representation with the corresponding number of digits.

■ **B:** The formula of the golden section $x/a = (a - x)/x$ (see 1.1.1.2, p. 2, 3.5.2.3, **3.**, p. 194 and 17.3.2.4, **4.**, p. 908) can be represented by the following two continued fractions: $x = a[1; 1, 1, \ldots]$ and $x = \frac{a}{2}(1 + \sqrt{5}) = \frac{a}{2}(1 + [2; 4, 4, 4, \ldots])$. The approximant α_4 delivers in the first case an accuracy of $0.018\,a$, in the second case of $0.000\,001\,a$.

1.1.1.5 Commensurability

Two numbers a and b are called *commensurable*, i.e., measurable by the same number, if both are an integer multiple of a third number c. From $a = mc$, $b = nc$ $(m, n \in \mathbb{Z})$ it follows that

$$\frac{a}{b} = x \;\; (x \text{ rational}). \tag{1.7}$$

Otherwise a and b are *incommensurable*.

■ **A:** The length of a side and a diagonal of a square are incommensurable because their ratio is the irrational number $\sqrt{2}$.

■ **B:** The lengths of the golden section are incommensurable, because their ratio contains the irrational number $\sqrt{5}$ (see 1.1.1.2, p. 2 and 3.5.2.3, **3.**,p. 194). Therefore the sides and diagonals in a regular pentagon are incommensurable (see ■ in 3.1.5.3, p. 139). Today Hippasos from Metapontum (450 BC) is considered to have discovered the irrational numbers via this example.

1.1.2 Methods for Proof

Mostly three types of proofs are used:
- direct proof,
- indirect proof,

• proof by (mathematical or arithmetical) induction.
Furthermore there are constructive proofs.

1.1.2.1 Direct Proof

The starting point is a theorem which has already been proven (premise p) and the truth of the statement of the new theorem is derived from it (conclusion q). The logical steps mostly used for the conclusions are implication and equivalence (see 5.1, p. 323).

1. Direct Proof by Implication

The *implication* $p \Rightarrow q$ means that the truth of the conclusion follows from the truth of the premise (see " Implication " in the truth table, 5.1.1, p. 323).

■ Prove the inequality $\frac{a+b}{2} \geq \sqrt{ab}$ for $a > 0,\ b > 0$. The premise is the well-known binomial formula $(a+b)^2 = a^2 + 2ab + b^2$. By subtracting $4ab$ follows $(a+b)^2 - 4ab = (a-b)^2 \geq 0$. From this inequality the statement is obtained certainly if the investigations are restricted only to the positive square roots because of $a > 0$ and $b > 0$.

2. Direct Proof by Equivalence

The proof will be delivered by *verifying* an equivalent statement. In practice it means that all the arithmetical operations which have to be used for changing p into q must be uniquely invertible.

■ Prove the inequality $1 + a + a^2 + \cdots + a^n < \frac{1}{1-a}$ for $0 < a < 1$.
Multiplying by $1-a$ yields $1 - a + a - a^2 + a^2 - a^3 \pm \cdots + a^n - a^{n+1} = 1 - a^{n+1} < 1$.
This last inequality is true because of the assumption $0 < a^{n+1} < 1$. The starting inequality also holds because all the arithmetical operations to be used are uniquely invertible.

1.1.2.2 Indirect Proof or Proof by Contradiction

To prove the statement q: Starting from its *negation* $\bar{q}$, and from $\bar{q}$ arriving at a false statement r, i.e., $\bar{q} \Rightarrow r$ (see also 5.1.1, **7.**, p. 325). In this case $\bar{q}$ must be false, because using the implication a false assumption can result only in a false conclusion (see truth table 5.1.1, p. 323). If $\bar{q}$ is false q must be true.

■ Prove that the number $\sqrt{2}$ is irrational. Suppose, $\sqrt{2}$ is rational. So the equality $\sqrt{2} = \frac{a}{b}$ holds for some integers a, b and $b \neq 0$. Assuming that the numbers a, b are *coprime numbers*, i.e., they do not have any common divisor, then follows $(\sqrt{2})^2 = 2 = \frac{a^2}{b^2}$ or $a^2 = 2b^2$, therefore, a^2 is an even number, and this is possible only if $a = 2n$ is an even number. Deducing $a^2 = 4n^2 = 2b^2$ holds, hence b must be an even number, too. It is obviously a contradiction to the assumption that a and b are coprimes.

1.1.2.3 Mathematical Induction

Theorems and dependent on natural numbers n are proven with this method. The principle of mathematical induction is the following: If the statement is valid for a natural number n_0, and if from the validity of the statement for a natural number $n \geq n_0$ the validity of the statement follows for $n+1$, then the statement is valid for every natural number $n \geq n_0$. According to these, the steps of the proof are:

1. Basis of the Induction: The truth of the statement is to be shown for $n = n_0$. Mostly $n_0 = 1$ can be choosen.

2. Induction Hypothesis: The statement is valid for an integer n (premise p).

3. Induction Conclusion: Formulation the proposition for $n+1$ (conclusion q).

4. Proof of the Implication: $p \Rightarrow q$.

Steps **3.** and **4.** together are called the *induction step* or *logical deduction from* n *to* $n+1$.

■ Prove the formula $s_n = \frac{1}{1\cdot 2} + \frac{1}{2\cdot 3} + \frac{1}{3\cdot 4} + \cdots + \frac{1}{n(n+1)} = \frac{n}{n+1}$.

The steps of the proof by induction are:

1. $n = 1:\ s_1 = \frac{1}{1\cdot 2} = \frac{1}{1+1}$ is obviously true.

2. Suppose $s_n = \frac{1}{1\cdot 2} + \frac{1}{2\cdot 3} + \frac{1}{3\cdot 4} + \cdots + \frac{1}{n(n+1)} = \frac{n}{n+1}$ holds for an $n \geq 1$.

3. Supposing **2.** it is to show: $s_{n+1} = \frac{n+1}{n+2}$.

4. The proof: $s_{n+1} = \frac{1}{1\cdot 2} + \frac{1}{2\cdot 3} + \frac{1}{3\cdot 4} + \cdots + \frac{1}{n(n+1)} + \frac{1}{(n+1)(n+2)} = s_n + \frac{1}{(n+1)(n+2)} =$
$\frac{n}{n+1} + \frac{1}{(n+1)(n+2)} = \frac{n^2+2n+1}{(n+1)(n+2)} = \frac{(n+1)^2}{(n+1)(n+2)} = \frac{n+1}{n+2}$.

1.1.2.4 Constructive Proof

In approximation theory, for instance, the proof of an existence theorem usually follows a *constructive* process, i.e., the steps of the proof give a method of calculation for a result which satisfies the propositions of the existence theorem.

■ The existence of a third-degree interpolation-spline function (see 19.7.1.1, **1.**, p. 996) can be proved in the following way: It is to be shown that the calculation of the coefficients of a spline satisfying the requirements of the existence theorem results in a tridiagonal linear equation system, which has a unique solution (see 19.7.1.1, **2.**, p. 997).

1.1.3 Sums and Products

1.1.3.1 Sums

1. Definition

To briefly denote a sum the *summation sign* $\sum$ is used:

$$a_1 + a_2 + \ldots + a_n = \sum_{k=1}^{n} a_k. \tag{1.8}$$

With this notation the sum of n summands a_k $(k = 1, 2, \ldots, n)$ is denoted, k is called the *running index* or *summation variable*.

2. Rules of Calculation

1. Sum of Summands Equal to Each Other , i.e., $a_k = a$ for $k = 1, 2, \ldots, n$:

$$\sum_{k=1}^{n} a_k = na. \tag{1.9a}$$

2. Multiplication by a Constant Factor

$$\sum_{k=1}^{n} ca_k = c\sum_{k=1}^{n} a_k. \tag{1.9b}$$

3. Separating a Sum

$$\sum_{k=1}^{n} a_k = \sum_{k=1}^{m} a_k + \sum_{k=m+1}^{n} a_k \quad (1 < m < n). \tag{1.9c}$$

4. Addition of Sums with the Same Length

$$\sum_{k=1}^{n}(a_k + b_k + c_k + \ldots) = \sum_{k=1}^{n} a_k + \sum_{k=1}^{n} b_k + \sum_{k=1}^{n} c_k + \ldots. \tag{1.9d}$$

5. Renumbering

$$\sum_{k=1}^{n} a_k = \sum_{k=m}^{m+n-1} a_{k-m+1}, \qquad \sum_{k=m}^{n} a_k = \sum_{k=l}^{n-m+l} a_{k+m-l}. \tag{1.9e}$$

6. Exchange the Order of Summation in Double Sums

$$\sum_{i=1}^{n} \left(\sum_{k=1}^{m} a_{ik} \right) = \sum_{k=1}^{m} \left(\sum_{i=1}^{n} a_{ik} \right). \tag{1.9f}$$

1.1.3.2 Products

1. Definition

The abbreviated notation for a product is the *product sign* $\prod$:

$$a_1 a_2 \ldots a_n = \prod_{k=1}^{n} a_k. \tag{1.10}$$

With this notation a product of n factors a_k $(k = 1, 2, \ldots, n)$ is denoted, where k is called the *running index*.

2. Rules of Calculation

1. Product of Coincident Factors , i.e., $a_k = a$ for $k = 1, 2, \ldots, n$:

$$\prod_{k=1}^{n} a_k = a^n. \tag{1.11a}$$

2. Factoring out a Constant Factor

$$\prod_{k=1}^{n} (c a_k) = c^n \prod_{k=1}^{n} a_k. \tag{1.11b}$$

3. Separating into Partial Products

$$\prod_{k=1}^{n} a_k = \left(\prod_{k=1}^{m} a_k \right) \left(\prod_{k=m+1}^{n} a_k \right) \quad (1 < m < n). \tag{1.11c}$$

4. Product of Products

$$\prod_{k=1}^{n} a_k b_k c_k \ldots = \left(\prod_{k=1}^{n} a_k \right) \left(\prod_{k=1}^{n} b_k \right) \left(\prod_{k=1}^{n} c_k \right) \ldots . \tag{1.11d}$$

5. Renumbering

$$\prod_{k=1}^{n} a_k = \prod_{k=m}^{m+n-1} a_{k-m+1}, \qquad \prod_{k=m}^{n} a_k = \prod_{k=l}^{n-m+l} a_{k+m-l}. \tag{1.11e}$$

6. Exchange the Order of Multiplication in Double Products

$$\prod_{i=1}^{n} \left(\prod_{k=1}^{m} a_{ik} \right) = \prod_{k=1}^{m} \left(\prod_{i=1}^{n} a_{ik} \right). \tag{1.11f}$$

1.1.4 Powers, Roots, and Logarithms

1.1.4.1 Powers

The notation a^x is used for the algebraic operation of *raising to a power*. The number a is called the *base*, x is called the *exponent* or *power*, and a^x is called the *power*. Powers are defined as in **Table 1.1**.

For the allowed values of bases and exponents there are the following
Rules of Calculation:

$$a^x\, a^y = a^{x+y}, \quad a^x : a^y = \frac{a^x}{a^y} = a^{x-y}, \tag{1.12}$$

$$a^x\, b^x = (a\, b)^x, \quad a^x : b^x = \frac{a^x}{b^x} = \left(\frac{a}{b}\right)^x, \tag{1.13}$$

$$(a^x)^y = (a^y)^x = a^{x\,y}, \tag{1.14}$$

$$a^x = e^{x \ln a} \quad (a > 0). \tag{1.15}$$

Here $\ln a$ is the natural logarithm of a where $e = 2.718281828459\ldots$ is the base. Special powers are

$$(-1)^n = \begin{cases} +1, & \text{if } n \text{ even}, \\ -1, & \text{if } n \text{ odd}, \end{cases} \quad (1.16a) \qquad a^0 = 1 \text{ for any } a \neq 0\,. \quad (1.16b)$$

Table 1.1 Definition of powers

base a	*exponent x*	*power* a^x
arbitrary real, $\neq 0$	0	1
	$n = 1, 2, 3, \ldots$	$a^n = \underbrace{a \cdot a \cdot a \cdot \ldots \cdot a}_{n \text{ factors}}$ (a to the power n)
	$n = -1, -2, -3, \ldots$	$a^n = \frac{1}{a^{-n}}$
positive real	rational: $\frac{p}{q}$ (p, q integer, $q > 0$)	$a^{\frac{p}{q}} = \sqrt[q]{a^p}$ (q-th root of a to the power p)
	irrational: $\lim\limits_{k\to\infty} \frac{p_k}{q_k}$	$\lim\limits_{k\to\infty} a^{\frac{p_k}{q_k}}$
0	positive	0

1.1.4.2 Roots

According to **Table 1.1** the n-th root of a positive number a is the positive number denoted by

$$\sqrt[n]{a} \quad (a > 0, \text{ real}; \; n > 0, \text{ integer}). \tag{1.17a}$$

This operation is called *taking of the root* or *extraction of the root*, a is the *radicand*, n is the *radical* or *index*.

The solution of the equation

$$x^n = a \quad (a \text{ real or complex}; \; n > 0\,, \text{ integer}) \tag{1.17b}$$

is often denoted by $x = \sqrt[n]{a}$. But there is no reason to be confused: In this relation the notation denotes all the solutions of the equation, i.e., it represents n different values x_k $(k = 1, 2, \ldots, n)$ to be calculated. In the cace of negative or complex values they are to be determined by (1.140b) (see 1.5.3.6, p. 38).

■ **A:** The equation $x^2 = 4$ has two real solutions, namely $x_{1,2} = \pm 2$.

■ **B:** The equation $x^3 = -8$ has three roots among the complex numbers: $x_1 = 1 + \mathrm{i}\sqrt{3}, x_2 = -2$ and $x_3 = 1 - \mathrm{i}\sqrt{3}$, but only one among the reals.

1.1.4.3 Logarithms

1. Definition

The *logarithm* u of a positive number $x > 0$ to the *base* $b > 0,\ b \neq 1$, is the exponent of the power which has the value x with b in the base. It is denoted by $u = \log_b x$. Consequently the equation

$$b^u = x \quad (1.18a) \qquad \text{yields} \qquad \log_b x = u \quad (1.18b)$$

and conversely the second one yields the first one. In particular holds

$$\log_b 1 = 0, \quad \log_b b = 1, \quad \log_b 0 = \begin{cases} -\infty \text{ for } b > 1, \\ +\infty \text{ for } b < 1. \end{cases} \qquad (1.18c)$$

The logarithm of negative numbers can be defined only among the complex numbers. The logarithmic functions see 2.6.2, p. 73.
To take the logarithm of a given number means to find its logarithm. To take the logarithm of an expression means it is transformed like (1.19a, 1.19b). The determination of a number or an expression from its logarithm is called *raising to a power*.

2. Some Properties of the Logarithm

a) Every positive number has a logarithm to any positive base, except the base $b = 1$.

b) For $x > 0$ and $y > 0$ the following **Rules of Calculation** are valid for any b (which is allowed to be a base):

$$\log(xy) = \log x + \log y, \quad \log\left(\frac{x}{y}\right) = \log x - \log y, \qquad (1.19a)$$

$$\log x^n = n \log x, \quad \text{in particular } \log \sqrt[n]{x} = \frac{1}{n} \log x\,. \qquad (1.19b)$$

With (1.19a, 1.19b) the logarithm of products and fractions can be calculated as sums or differences of logarithms .

■ Take the logarithm of the expression $\dfrac{3x^2\sqrt[3]{y}}{2zu^3}$: $\log \dfrac{3x^2\sqrt[3]{y}}{2zu^3} = \log\left(3x^2\sqrt[3]{y}\right) - \log(2zu^3)$
$= \log 3 + 2\log x + \dfrac{1}{3}\log y - \log 2 - \log z - 3\log u.$

Often the reverse transformation is required, i.e., an expression containing logarithms of different amounts is to be rewritten into one, which is the logarithm of one expression.

■ $\log 3 + 2\log x + \dfrac{1}{3}\log y - \log 2 - \log z - 3\log u = \log \dfrac{3x^2\sqrt[3]{y}}{2zu^3}.$

c) Logarithms to different bases are proportional, i.e., the logarithm to a base a can be change into a logarithm to the base b by multiplication:

$$\log_a x = M \log_b x \ \text{ where } \ M = \log_a b = \frac{1}{\log_b a}. \qquad (1.20)$$

M is called the *modulus of the transformation*.

1.1.4.4 Special Logarithms

1. The logarithm to the base 10 is called the *decimal* or *Briggsian logarithm*, in formulas:

$$\log_{10} x = \lg x \ \text{ and } \ \log(x 10^\alpha) = \alpha + \log x. \qquad (1.21)$$

2. The logarithm to the base e is called the *natural* or *Neperian logarithm*, in formulas:

$$\log_e x = \ln x. \qquad (1.22)$$

The modulus of transformation to change from the natural logarithm into the decimal one is

$$M = \log e = \frac{1}{\ln 10} = 0.4342944819\,, \qquad (1.23)$$

and to change from the decimal into the natural one it is

$$M_1 = \frac{1}{M} = \ln 10 = 2.3025850930\,. \tag{1.24}$$

3. The logarithm to base 2 is called the *binary logarithm*, in formulas:

$$\log_2 x = \operatorname{ld} x \quad \text{or} \quad \log_2 x = \operatorname{lb} x. \tag{1.25}$$

4. The values of the decimal and natural logarithm can be found in *logarithm tables*. Some time ago the logarithm was used for numerical calculation of powers, and it often made numerical multiplication and division easier. Mostly the decimal logarithm was used. Today pocket calculators and personal computers make these calculations.
Every number given in decimal form (so every real number), which is called in this relation the *antilog*, can be written in the form

$$x = \hat{x}10^k \text{ with } 1 \le \hat{x} < 10 \tag{1.26a}$$

by factoring out an appropriate power of ten: 10^k with integer k. This form is called the *half-logarithmic representation*. Here $\hat{x}$ is given by the sequence of figures of x, and 10^k is the order of magnitude of x. Then for the logarithm holds

$$\log x = k + \log \hat{x} \quad \text{with} \quad 0 \le \log \hat{x} < 1, \quad \text{i.e.,} \quad \log \hat{x} = 0,\ldots\,. \tag{1.26b}$$

Here k is the so-called *characteristic* and the sequence of figures behind the decimal point of $\log \hat{x}$ is called the *mantissa*. The mantissa can be found in logarithm tables.

■ $\lg 324 = 2.5105$, the characteristic is 2, the mantissa is 5105. Multiplying or dividing this number by 10^n, for example 324000; 3240; 3.24; 0.0324, their logarithms have the same mantissa, here 5105, but different characteristics. That is why the mantissas are given in *logarithm tables*. In order to get the mantissa of a number x first the decimal point has to be moved to the right or to the left to get a number between 1 and 10, and the characteristic of the antilog x is determined by how many digits k the decimal point was moved.

5. Slide rule Beside the logarithm, the *slide rule* was of important practical help in numerical calculations. The slide rule works by the principle of the form (1.19a), so multiplying and dividing is done by adding and subtracting numbers. On the slide rule the scale-segments are denoted according to the logarithm values, so multiplication and division can be performed as addition or subtraction (see Scale and Graph Papers 2.17.1, p. 115).

1.1.5 Algebraic Expressions

1.1.5.1 Definitions

1. Algebraic Expression

One or more algebraic quantities, such as numbers or symbols, are called an *algebraic expression* or *term* if they are connected by the symbols, $+$, $-$, $\cdot$, $:$, $\sqrt{}$, etc., as well as by different types of braces for fixing the order of operations.

2. Identity

is an equality relation between two algebraic expressions if for arbitrary values of the symbols in them the equality holds.

3. Equation

is an equality relation between two algebraic expressions if the equality holds only for a few values of the symbols. For instance an equality relation

$$F(x) = f(x) \tag{1.27}$$

between two functions with the same independent variable is considered as an *equation with one variable* if it holds only for certain values of the variable. If the equality is valid for every value of x, it is called an identity, or one says the equality holds identically, written as formula $F(x) \equiv f(x)$.

4. Identical Transformations
are performed in order to change an algebraic expression into another one if the two expression are identically equal. The goal is to have another form, e.g., to get a shorter form or a more convenient form for further calculations. Often it is of interest to have the expression in a form which is especially good for solving an equation, or taking the logarithm, or calculating the derivative or integral of it, etc.

1.1.5.2 Algebraic Expressions in Detail

1. Principal Quantities
Principal quantities are those general numbers (literal symbols) occurring in algebraic expressions, according to which the expressions are classified. They must be fixed in any single case. In the case of functions, the independent variables are the principal quantities. The other quantities not given by numbers are the *parameters* of the expression. In some expressions the parameters are called *coefficients*.

■ So-called coefficients occur e.g. in the cases of polynomials, Fourier series, and linear differential equations, etc.

An expression belongs to a certain class depending on which kind of operations are performed on the principal quantities. Usually, the last letters of the alphabet $x, y, z, u, v, \ldots$ are used to denote the principal quantities and the first letters $a, b, c, \ldots$ are used for parameters. The letters $m, n, p, \ldots$ are usually used for positive integer parameter values, e.g. for indices in summations or in iterations.

2. Integral Rational Expressions
are expressions which contain only addition, subtraction, and multiplication of the principal quantities, including powers of them with non-negative integer exponents.

3. Rational Expressions
contain also division by principal quantities, i.e., division by integral rational expressions, so principal quantities can have negative integers in the exponent.

4. Irrational Expressions
contain roots, i.e., non-integer rational powers of integral rational or rational expressions with respect to their principal quantities, of course.

5. Transcendental Expressions
contain exponential, logarithmic or trigonometric expressions of the principal quantities, i.e., there can be irrational numbers in the exponent of an expression of principal quantities, or an expression of principal quantities can be in the exponent, or in the argument of a trigonometric or logarithmic expression.

1.1.6 Integral Rational Expressions

1.1.6.1 Representation in Polynomial Form

Every integral rational expression can be changed into polynomial form by elementary transformations, as in addition, subtraction, and multiplication of monomials and polynomials.

■ $(-a^3 + 2a^2x - x^3)(4a^2 + 8ax) + (a^3x^2 + 2a^2x^3 - 4ax^4) - (a^5 + 4a^3x^2 - 4ax^4)$
$= -4a^5 + 8a^4x - 4a^2x^3 - 8a^4x + 16a^3x^2 - 8ax^4 + a^3x^2 + 2a^2x^3 - 4ax^4 - a^5 - 4a^3x^2 + 4ax^4$
$= -5a^5 + 13a^3x^2 - 2a^2x^3 - 8ax^4.$

1.1.6.2 Factoring Polynomials

Polynomials often can be decomposed into a product of monomials and polynomials. To do so, *factoring out*, *grouping*, special formulas and special properties of equations can be used.

■ **A:** Factoring out: $8ax^2y - 6bx^3y^2 + 4cx^5 = 2x^2(4ay - 3bxy^2 + 2cx^3)$.

■ **B:** Grouping: $6x^2 + xy - y^2 - 10xz - 5yz = 6x^2 + 3xy - 2xy - y^2 - 10xz - 5yz = 3x(2x + y) - y(2x + y) - 5z(2x + y) = (2x + y)(3x - y - 5z)$.

■ **C:** Using the properties of equations (see also 1.6.3.1, p. 43): $P(x) = x^6 - 2x^5 + 4x^4 + 2x^3 - 5x^2$. **a)** Factoring out x^2. **b)** Realizing that $\alpha_1 = 1$ and $\alpha_2 = -1$ are the roots of the equation $P(x) = 0$ and dividing $P(x)$ by $x^2(x-1)(x+1) = x^4 - x^2$ gives the quotient $x^2 - 2x + 5$. This expression can no longer be decomposed into real factors because $p = -2$, $q = 5$, $p^2/4 - q < 0$, so finally the decomposition is $x^6 - 2x^5 + 4x^4 + 2x^3 - 5x^2 = x^2(x-1)(x+1)(x^2 - 2x + 5)$.

1.1.6.3 Special Formulas

$$(x \pm y)^2 = x^2 \pm 2xy + y^2, \tag{1.28}$$

$$(x + y + z)^2 = x^2 + y^2 + z^2 + 2xy + 2xz + 2yz, \tag{1.29}$$

$$(x + y + z + \cdots + t + u)^2 = x^2 + y^2 + z^2 + \cdots + t^2 + u^2 + \\ +2xy + 2xz + \cdots + 2xu + 2yz + \cdots + 2yu + \cdots + 2tu, \tag{1.30}$$

$$(x \pm y)^3 = x^3 \pm 3x^2y + 3xy^2 \pm y^3. \tag{1.31}$$

The calculation of the expression $(x \pm y)^n$ is done by the help of the binomial formula (see (1.36a)–(1.37a)).

$$(x + y)(x - y) = x^2 - y^2, \tag{1.32}$$

$$\frac{x^n - y^n}{x - y} = x^{n-1} + x^{n-2}y + \cdots + xy^{n-2} + y^{n-1}, \quad \text{(for integer } n\text{, and } n > 1), \tag{1.33}$$

$$\frac{x^n + y^n}{x + y} = x^{n-1} - x^{n-2}y + \cdots - xy^{n-2} + y^{n-1} \quad \text{(for odd } n\text{, and } n > 1), \tag{1.34}$$

$$\frac{x^n - y^n}{x + y} = x^{n-1} - x^{n-2}y + \cdots + xy^{n-2} - y^{n-1} \quad \text{(for even } n\text{, and } n > 1). \tag{1.35}$$

1.1.6.4 Binomial Theorem

1. Power of an Algebraic Sum of Two Summands (First Binomial Formula)

The formula

$$(a + b)^n = a^n + na^{n-1}b + \frac{n(n-1)}{2!}a^{n-2}b^2 + \frac{n(n-1)(n-2)}{3!}a^{n-3}b^3 \\ + \cdots + \frac{n(n-1)\ldots(n-m+1)}{m!}a^{n-m}b^m + \cdots + nab^{n-1} + b^n \tag{1.36a}$$

is called the *binomial theorem*, where a and b are real or complex values and $n = 1, 2, \ldots$. Using the *binomial coefficients* delivers a shorter and more convenient notation:

$$(a + b)^n = \binom{n}{0}a^n + \binom{n}{1}a^{n-1}b + \binom{n}{2}a^{n-2}b^2 + \binom{n}{3}a^{n-3}b^3 + \cdots + \binom{n}{n-1}ab^{n-1} + \binom{n}{n}b^n \tag{1.36b}$$

or

$$(a + b)^n = \sum_{k=0}^{n} \binom{n}{k} a^{n-k}b^k. \tag{1.36c}$$

2. Power of an Algebraic Difference (Second Binomial Formula)

$$(a - b)^n = a^n - na^{n-1}b + \frac{n(n-1)}{2!}a^{n-2}b^2 - \frac{n(n-1)(n-2)}{3!}a^{n-3}b^3 \\ + \cdots + (-1)^m\frac{n(n-1)\ldots(n-m+1)}{m!}a^{n-m}b^m + \cdots + (-1)^n b^n \tag{1.37a}$$

or

$$(a-b)^n = \sum_{k=0}^{n} \binom{n}{k} (-1)^k a^{n-k} b^k. \tag{1.37b}$$

3. Binomial Coefficients

The definition is for non-negative and integer n and k:

$$\binom{n}{k} = \frac{n!}{(n-k)!k!} \quad (0 \le k \le n), \tag{1.38a}$$

where $n!$ is the product of the positive integers from 1 to n, and it is called n *factorial*:

$$n! = 1 \cdot 2 \cdot 3 \cdot \ldots \cdot n, \text{ and by definition } 0! = 1. \tag{1.38b}$$

The binomial coefficients can easily be seen from the *Pascal triangle* in **Table 1.2**. The first and the last number is equal to one in every row; every other coefficient is the sum of the numbers standing on left and on right in the row above it.

Simple calculations verify the following formulas:

$$\binom{n}{k} = \binom{n}{n-k} = \frac{n!}{k!(n-k)!}, \tag{1.39a}$$

$$\binom{n}{0} = 1, \quad \binom{n}{1} = n, \quad \binom{n}{n} = 1. \tag{1.39b}$$

$$\binom{n+1}{k+1} = \binom{n}{k} + \binom{n-1}{k} + \binom{n-2}{k} + \cdots + \binom{k}{k}. \tag{1.39c}$$

$$\binom{n+1}{k} = \frac{n+1}{n-k+1}\binom{n}{k}. \tag{1.39d}$$

$$\binom{n}{k+1} = \frac{n-k}{k+1}\binom{n}{k}. \tag{1.39e}$$

$$\binom{n+1}{k+1} = \binom{n}{k+1} + \binom{n}{k}. \tag{1.39f}$$

Table 1.2 Pascal's triangle

n	Coefficients												
0							1						
1						1		1					
2					1		2		1				
3				1		3		3		1			
4			1		4		6		4		1		
5		1		5		10		10		5		1	
6	1		6		15		20		15		6		1
	↑		↑		↑		↑		↑		↑		↑
6	$\binom{6}{0}$		$\binom{6}{1}$		$\binom{6}{2}$		$\binom{6}{3}$		$\binom{6}{4}$		$\binom{6}{5}$		$\binom{6}{6}$
⋮	. . .	.	.	.	.	.	.	.	.	.	.	.	. . .

For an arbitrary real value α ($\alpha \in \mathbb{R}$) and a non-negative integer k one can define the binomial coefficient $\binom{\alpha}{k}$:

$$\binom{\alpha}{k} = \frac{\alpha(\alpha-1)(\alpha-2)\cdots(\alpha-k+1)}{k!} \quad \text{for integer } k \text{ and } k \ge 1, \quad \binom{\alpha}{0} = 1. \tag{1.40}$$

■ $\binom{-\frac{1}{2}}{3} = \frac{-\frac{1}{2}(-\frac{1}{2}-1)(-\frac{1}{2}-2)}{3!} = -\frac{5}{16}$.

4. Properties of the Binomial Coefficients

- The binomial coefficients increase until the middle of the binomial formula (1.36b), then decrease.
- The binomial coefficients are equal for the terms standing in symmetric positions with respect to the start and the end of the expression.
- The sum of the binomial coefficients in the binomial formula of degree n is equal to 2^n .
- The sum of the coefficients at the odd positions is equal to the sum of the coefficients at the even positions.

5. Binomial Series

The formula (1.36a) of the binomial theorem can also be extended for negative and fraction exponents. If $|b| < a$, then $(a+b)^n$ has a *convergent infinite series* (see also 21.5, p. 1057):

$$(a+b)^n = a^n + na^{n-1}b + \frac{n(n-1)}{2!}a^{n-2}b^2 + \frac{n(n-1)(n-2)}{3!}a^{n-3}b^3 + \cdots . \tag{1.41}$$

1.1.6.5 Determination of the Greatest Common Divisor of Two Polynomials

It is possible that two polynomials $P(x)$ of degree n and $Q(x)$ of degree m with $n \geq m$ have a common polynomial factor, which contains x. The least common multiple of these factors is the *greatest common divisor* of the polynomials.

■ $P(x) = (x-1)^2(x-2)(x-4)$, $Q(x) = (x-1)(x-2)(x-3)$; the greates common devisor is $(x-1)(x-2)$.

If $P(x)$ and $Q(x)$ do not have any common polynomial factor, they are called *relatively prime* or *coprime*. In this case, their greatest common divisor is a constant.

The greatest common divisor of two polynomials $P(x)$ and $Q(x)$ can be determined by the *Euclidean algorithm* without decomposing them into factors:

1. Division of $P(x)$ by $Q(x) = R_0(x)$ results in the quotient $T_1(x)$ and the remainder $R_1(x)$:

$$P(x) = Q(x)T_1(x) + R_1(x)\,. \tag{1.42a}$$

2. Division of $Q(x)$ by $R_1(x)$ results in the quotient $T_2(x)$ and the remainder $R_2(x)$:

$$Q(x) = R_1(x)T_2(x) + R_2(x). \tag{1.42b}$$

3. Division of $R_1(x)$ by $R_2(x)$ results in $T_3(x)$ and $R_3(x)$, etc. The greatest common divisor of the two polynomials is the last non-zero remainder $R_k(x)$. This method is known from the arithmetic of natural numbers (see 1.1.1.4, p. 3).

The determination of the greatest common divisor can be used, e. g., when equations musz be solved to separate the roots with higher multiplicity or to apply the Sturm method (see 1.6.3.2, **2.**, p. 44).

1.1.7 Rational Expressions

1.1.7.1 Reducing to the Simplest Form

Every rational expression can be written in the form of a quotient of two coprime polynomials. To do this, only elementary transformations are necessary such as addition, subtraction, multiplication and division of polynomials and fractions and simplification of fractions.

■ Find the most simple form of $\dfrac{3x + \dfrac{2x+y}{z}}{x\left(x^2 + \dfrac{1}{z^2}\right)} - y^2 + \dfrac{x+z}{z}$:

$$\frac{(3xz+2x+y)z^2}{(x^3z^2+x)z}+\frac{-y^2z+x+z}{z}=\frac{3xz^3+2xz^2+yz^2+(x^3z^2+x)(-y^2z+x+z)}{x^3z^3+xz}=$$
$$\frac{3xz^3+2xz^2+yz^2-x^3y^2z^3-xy^2z+x^4z^2+x^2+x^3z^3+xz}{x^3z^3+xz}\,.$$

1.1.7.2 Determination of the Integral Rational Part

A quotient of two polynomials with the same variable x is a *proper fraction* if the degree of the numerator is less than the degree of the denominator. In the opposite case, it is called an *improper fraction.* Every improper fraction can be decomposed into a sum of a proper fraction and a polynomial by dividing the numerator by the denominator, i.e., separating the integral rational part.

■ Determine the integral rational part of $R(x)=\dfrac{3x^4-10ax^3+22a^2x^2-24a^3x+10a^4}{x^2-2ax+3a^2}$:

$$(3x^4-10ax^3+22a^2x^2-24a^3x+10a^4):(x^2-2ax+3a^2)=3x^2-4ax+5a^2+\frac{-2a^3x-5a^4}{x^2-2ax-3a^2}$$

$$\begin{array}{l}
3x^4-\ 6ax^3+\ 9a^2x^2\\
\hline
\quad -\ 4ax^3+13a^2x^2-24a^3x\\
\quad -\ 4ax^3+\ 8a^2x^2-12a^3x\\
\hline
\qquad\qquad 5a^2x^2-12a^3x+10a^4\\
\qquad\qquad 5a^2x^2-10a^3x+15a^4\\
\hline
\qquad\qquad\quad -\ 2a^3x-\ 5a^4.
\end{array}$$

$$R(x)=3x^2-4ax+5a^2+\frac{-2a^3x-5a^4}{x^2-2ax+3a^2}.$$

The integral rational part of a rational function $R(x)$ is considered to be as an *asymptotic approximation* for $R(x)$ because for large values of $|x|$, the value of the proper fraction part tends to zero, and $R(x)$ behaves as its polynomial part.

1.1.7.3 Partial Fraction Decomposition

Every proper rational fraction

$$R(x)=\frac{P(x)}{Q(x)}=\frac{a_nx^n+a_{n-1}x^{n-1}+\cdots+a_1x+a_0}{b_mx^m+b_{m-1}x^{m-1}+\cdots+b_1x+b_0}\quad(n<m)\tag{1.43}$$

with coprime polynomials in the numerator and denominator can be uniquely decomposed into a sum of partial fractions. The coefficients $a_0,a_1,\ldots,a_n,b_0,b_1,\ldots,b_n$ are real or complex numbers. The partial fractions have the form

$$\frac{A}{(x-\alpha)^k}\quad(1.44a)\quad\text{and}\quad\frac{Dx+E}{(x^2+px+q)^m}\quad\text{where}\quad\left(\frac{p}{2}\right)^2-q<0\,.\quad(1.44b)$$

In the followings real coefficients are assumed in $R(x)$ in (1.43).
First the leading coefficient b_m of the denominator $Q(x)$ is transformed into 1 by dividing the numerator and the denominator of (1.43) by the original value of b_m. In the case of real coefficients the following three cases are to be distinguished.
In the case of complex coefficients in $R(x)$ only the first two cases can occur, since complex polynomials can be factorized into a product of first degree polynomials. Every proper rational fraction $R(x)$ can be expanded into a sum of fractions of the form (1.44a), where A and α are complex numbers.

1. Partial Fraction Decomposition, Case 1

The denominator $Q(x)$ has m different simple roots $\alpha_1,\ldots,\alpha_m$. Then the expansion has the form

$$\frac{P(x)}{Q(x)}=\frac{a_nx^n+\cdots+a_0}{(x-\alpha_1)(x-\alpha_2)\ldots(x-\alpha_m)}=\frac{A_1}{x-\alpha_1}+\frac{A_2}{x-\alpha_2}+\cdots+\frac{A_m}{x-\alpha_m}\tag{1.45a}$$

with coefficients

$$A_1 = \frac{P(\alpha_1)}{Q'(\alpha_1)}, \quad A_2 = \frac{P(\alpha_2)}{Q'(\alpha_2)}, \quad \ldots, \quad A_m = \frac{P(\alpha_m)}{Q'(\alpha_m)}, \tag{1.45b}$$

where in the numerators of (1.45b) the values of the derivative $\frac{dQ}{dx}$ are taken for $x = \alpha_1,\ x = \alpha_2, \ldots$.

■ $\frac{6x^2 - x + 1}{x^3 - x} = \frac{A}{x} + \frac{B}{x-1} + \frac{C}{x+1}$, $\alpha_1 = 0$, $\alpha_2 = +1$ and $\alpha_3 = -1$;
$P(x) = 6x^2 - x + 1$, $Q'(x) = 3x^2 - 1$, $A = \frac{P(0)}{Q'(0)} = -1$, $B = \frac{P(1)}{Q'(1)} = 3$ and $C = \frac{P(-1)}{Q'(-1)} = 4$,
$\frac{P(x)}{Q(x)} = -\frac{1}{x} + \frac{3}{x-1} + \frac{4}{x+1}$.

An other possibility to determine the coefficients $A_1,\ A_2, \ldots, A_m$ is the method of comparing coefficients (see **4.**, p. 17).

2. Partial Fraction Decomposition, Case 2

The denominator $Q(x)$ has l multiple real roots $\alpha_1, \alpha_2, \ldots, \alpha_l$ with multiplicities $k_1, k_2, \ldots, k_l$ respectively. Then the decomposition has the form

$$\begin{aligned}\frac{P(x)}{Q(x)} &= \frac{a_n x^n + a_{n-1}x^{n-1} + \cdots + a_0}{(x-\alpha_1)^{k_1}(x-\alpha_2)^{k_2}\ldots(x-\alpha_l)^{k_l}} = \frac{A_1}{x-\alpha_1} + \frac{A_2}{(x-\alpha_1)^2} + \cdots + \frac{A_{k_1}}{(x-\alpha_1)^{k_1}} \\ &\quad + \frac{B_1}{x-\alpha_2} + \frac{B_2}{(x-\alpha_2)^2} + \cdots + \frac{B_{k_2}}{(x-\alpha_2)^{k_2}} + \cdots + \frac{L_{k_l}}{(x-\alpha_l)^{k_l}}.\end{aligned} \tag{1.46}$$

■ $\frac{x+1}{x(x-1)^3} = \frac{A_1}{x} + \frac{B_1}{x-1} + \frac{B_2}{(x-1)^2} + \frac{B_3}{(x-1)^3}$. The coefficients A_1, B_1, B_2, B_3 can be determined by the method of comparing coefficients.

3. Partial Fraction Decomposition, Case 3

If the denominator $Q(x)$ has also complex roots, then its factorization is

$$\begin{aligned}Q(x) &= (x-\alpha_1)^{k_1}(x-\alpha_2)^{k_2}\cdots(x-\alpha_l)^{k_l} \\ &\quad \cdot(x^2 + p_1x + q_1)^{m_1}(x^2 + 2p_2x + q_2)^{m_2}\cdots(x^2 + p_rx + q_r)^{m_r}\end{aligned} \tag{1.47}$$

according to (1.168), p. 44. Here $\alpha_1, \alpha_2, \ldots, \alpha_l$ are the l real roots of polynomial $Q(x)$. Beside these roots $Q(x)$ has r complex conjugate pairs of roots, which are the roots of the quadratic factors $x^2 - p_ix + q_i$ (i =1,2,...,r). The numbers p_i, q_i are real, and $\left(\frac{p_i}{2}\right)^2 - q_i < 0$ holds. In this case the partial fraction decomposition has the form

$$\begin{aligned}\frac{P(x)}{Q(x)} &= \frac{a_nx^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0}{(x-\alpha_1)^{k_1}(x-\alpha_2)^{k_2}\cdots(x^2+p_1x+q_1)^{m_1}(x^2+p_2x+q_2)^{m_2}\cdots} \\ &= \frac{A_1}{x-\alpha_1} + \frac{A_2}{(x-\alpha_1)^2} + \cdots + \frac{A_{k_1}}{(x-\alpha_1)^{k_1}} + \frac{B_1}{x-\alpha_2} + \frac{B_2}{(x-\alpha_2)^2} + \cdots + \frac{B_{k_2}}{(x-\alpha_2)^{k_2}} + \cdots \\ &\quad + \frac{C_1x + D_1}{x^2 + p_1x + q_1} + \frac{C_2x + D_2}{(x^2 + p_1x + q_1)^2} + \cdots + \frac{C_{m_1}x + D_{m_1}}{(x^2 + p_1x + q_1)^{m_1}} + \\ &\quad + \frac{E_1x + F_1}{x^2 + p_2x + q_2} + \frac{E_2x + F_2}{(x^2 + p_2x + q_2)^2} + \cdots + \frac{E_{m_2}x + F_{m_2}}{(x^2 + p_2x + q_2)^{m_2}} + \cdots.\end{aligned} \tag{1.48}$$

■ $\frac{5x^2 - 4x + 16}{(x-3)(x^2-x+1)^2} = \frac{A}{x-3} + \frac{C_1x + D_1}{x^2 - x + 1} + \frac{C_2x + D_2}{(x^2-x+1)^2}$. The coefficients A, C_1, D_1, C_2, D_2 are to be determined by the method of comparing coefficients.

4. Method of Comparing Coefficients

In order to determine the coefficients $A_1, A_2, \ldots, E_1, F_1 \ldots$ in (1.48) the expression (1.48) has to be multiplied by $Q(x)$, then the result $Z(x)$ is compared with $P(x)$, since $Z(x) \equiv P(x)$. After ordering $Z(x)$ by the powers of x, one gets a system of equations by comparing the coefficients of the corresponding x–powers in $Z(x)$ and $P(x)$. This method is called the *method of comparing coefficients* or *method of undetermined coefficients*.

■ $\dfrac{6x^2 - x + 1}{x^3 - x} = \dfrac{A}{x} + \dfrac{B}{x-1} + \dfrac{C}{x+1} = \dfrac{A(x^2-1) + Bx(x+1) + Cx(x-1)}{x(x^2-1)}$.

Comparing the coefficients of the same powers of x, one gets the system of equations $6 = A + B + C$, $-1 = B - C$, $1 = -A$, and its solutions are $A = -1, B = 3, C = 4$.

1.1.7.4 Transformations of Proportions

The equality

$$\frac{a}{b} = \frac{c}{d} \quad (1.49a) \qquad \text{yields} \qquad ad = bc, \quad \frac{a}{c} = \frac{b}{d}, \quad \frac{d}{b} = \frac{c}{a}, \quad \frac{b}{a} = \frac{d}{c} \quad (1.49b)$$

and furthermore

$$\frac{a \pm b}{b} = \frac{c \pm d}{d}, \quad \frac{a \pm b}{a} = \frac{c \pm d}{c}, \quad \frac{a \pm c}{c} = \frac{b \pm d}{d}, \quad \frac{a+b}{a-b} = \frac{c+d}{c-d}. \quad (1.49c)$$

From the equalities of the proportions

$$\frac{a_1}{b_1} = \frac{a_2}{b_2} = \cdots = \frac{a_n}{b_n} \quad (1.50a) \qquad \text{it follows that} \qquad \frac{a_1 + a_2 + \cdots + a_n}{b_1 + b_2 + \cdots + b_n} = \frac{a_1}{b_1}. \quad (1.50b)$$

1.1.8 Irrational Expressions

Every irrational expression can be written in a simpler form by 1. simplifying the exponent, 2. taking out terms from the radical sign and 3. moving the irrationality into the numerator.

1. Simplifying the Exponent The exponent can be simplified if the radicand can be factorized and the index of the radical and the exponents in the radicand have a common factor; the index of the radical and the exponents must be divided by their greatest common divisor.

■ $\sqrt[6]{16(x^{12} - 2x^{11} + x^{10})} = \sqrt[6]{4^2 \cdot x^{5\cdot 2}(x-1)^2} = \sqrt[3]{4x^5(x-1)}$.

2. Moving the Irrationality There are different ways to *move the irrationality* into the numerator.

■ **A:** $\sqrt{\dfrac{x}{2y}} = \sqrt{\dfrac{2xy}{4y^2}} = \dfrac{\sqrt{2xy}}{2y}$. ■ **B:** $\sqrt[3]{\dfrac{x}{4yz^2}} = \sqrt[3]{\dfrac{2xy^2z}{8y^3z^3}} = \dfrac{\sqrt[3]{2xy^2z}}{2yz}$.

■ **C:** $\dfrac{1}{x + \sqrt{y}} = \dfrac{x - \sqrt{y}}{\left(x + \sqrt{y}\right)\left(x - \sqrt{y}\right)} = \dfrac{x - \sqrt{y}}{x^2 - y}$.

■ **D:** $\dfrac{1}{x + \sqrt[3]{y}} = \dfrac{x^2 - x\sqrt[3]{y} + \sqrt[3]{y^2}}{\left(x + \sqrt[3]{y}\right)\left(x^2 - x\sqrt[3]{y} + \sqrt[3]{y^2}\right)} = \dfrac{x^2 - x\sqrt[3]{y} + \sqrt[3]{y^2}}{x^3 + y}$.

3. Simplest Forms of Powers and Radicals Also powers and radicals can be transformed into the simplest form.

■ **A:** $\sqrt[4]{\dfrac{81x^6}{(\sqrt{2} - \sqrt{x})^4}} = \sqrt{\dfrac{9x^3}{(\sqrt{2} - \sqrt{x})^2}} = \dfrac{3x\sqrt{x}}{\sqrt{2} - \sqrt{x}} = \dfrac{3x\sqrt{x}(\sqrt{2} + \sqrt{x})}{2 - x} = \dfrac{3x\sqrt{2x} + 3x^2}{2 - x}$.

■ **B:** $\left(\sqrt{x} + \sqrt[3]{x^2} + \sqrt[4]{x^3} + \sqrt[12]{x^7}\right)\left(\sqrt{x} - \sqrt[3]{x} + \sqrt[4]{x} - \sqrt[12]{x^5}\right) = (x^{1/2} + x^{2/3} + x^{3/4} + x^{7/12})(x^{1/2} - x^{1/3} + x^{1/4} - x^{5/12}) = x + x^{7/6} + x^{5/4} + x^{13/12} - x^{5/6} - x - x^{13/12} - x^{11/12} + x^{3/4} + x^{11/12} + x + x^{5/6} - x^{11/12} - x^{13/12} -$

$x^{7/6} - x = x^{5/4} - x^{13/12} - x^{11/12} + x^{3/4} = \sqrt[4]{x^5} - \sqrt[12]{x^{13}} - \sqrt[12]{x^{11}} + \sqrt[4]{x^3} = x^{3/4}(1 - x^{1/6} - x^{1/3} + x^{1/2}) = \sqrt[4]{x^3}\,(1 - \sqrt[6]{x} - \sqrt[3]{x} + \sqrt{x})$.

1.2 Finite Series

1.2.1 Definition of a Finite Series

The sum

$$s_n = a_0 + a_1 + a_2 + \cdots + a_n = \sum_{i=0}^{n} a_i, \tag{1.51}$$

is called a *finite series*. The summands a_i $(i = 0, 1, 2, \ldots, n)$ are given by certain formulas, they are numbers, and they are the *terms of the series*.

1.2.2 Arithmetic Series

1. Arithmetic Series of First Order

is a finite series where the terms form an *arithmetic sequence*, i.e., the difference of two terms standing after each other is a constant:

$$\Delta a_i = a_{i+1} - a_i = d = \text{const} \quad \text{holds,} \quad \text{so} \quad a_i = a_0 + id. \tag{1.52a}$$

Thus holds:

$$s_n = a_0 + (a_0 + d) + (a_0 + 2d) + \cdots + (a_0 + nd) \tag{1.52b}$$

$$s_n = \frac{a_0 + a_n}{2}(n+1) = \frac{n+1}{2}(2a_0 + nd). \tag{1.52c}$$

2. Arithmetic Series of k-th Order

is a finite series, where the k-th differences $\Delta^k a_i$ of the sequence a_0, a_1, a_2, $\ldots$, a_n are constants. The differences of higher order are calculated by the formula

$$\Delta^\nu a_i = \Delta^{\nu-1} a_{i+1} - \Delta^{\nu-1} a_i \quad (\nu = 2, 3, \ldots, k). \tag{1.53a}$$

It is convenient to calculate them from the *difference schema* (also difference table or triangle schema):

$$\begin{array}{lllllll}
a_0 & & & & & & \\
 & \Delta a_0 & & & & & \\
a_1 & & \Delta^2 a_0 & & & & \\
 & \Delta a_1 & & \Delta^3 a_0 & & & \\
a_2 & & \Delta^2 a_1 & & \ddots\ \Delta^k a_0 & & \\
 & \Delta a_2 & & \Delta^3 a_1 & & & \\
a_3 & & \Delta^2 a_2 & & \ddots\ \Delta^k a_1 & \ddots & \\
 & \vdots & & \vdots & \vdots & & \Delta^n a_0 \\
\vdots & & \vdots & & \Delta^k a_{n-k} & \iddots & \\
 & & & \Delta^3 a_{n-3}\ \iddots & & & \\
 & & \Delta^2 a_{n-2} & & & & \\
 & \Delta a_{n-1} & & & & & \\
a_n & & & & & &
\end{array} \tag{1.53b}$$

The following formulas hold for the terms and the sum:

$$a_i = a_0 + \binom{i}{1}\Delta a_0 + \binom{i}{2}\Delta^2 a_0 + \cdots + \binom{i}{k}\Delta^k a_0 \quad (i = 1, 2, \ldots, n), \tag{1.53c}$$

$$s_n = \binom{n+1}{1} a_0 + \binom{n+1}{2}\Delta a_0 + \binom{n+1}{3}\Delta^2 a_0 + \cdots + \binom{n+1}{k+1}\Delta^k a_0. \tag{1.53d}$$

1.2.3 Geometric Series

The sum (1.51) is called a *geometric series*, if the terms form a *geometric sequence*, i.e., the ratio of two successive terms is a constant:

$$\frac{a_{i+1}}{a_i} = q = \text{const} \quad \text{holds,} \quad \text{so} \quad a_i = a_0 q^i. \tag{1.54a}$$

Thus holds:

$$s_n = a_0 + a_0 q + a_0 q^2 + \cdots + a_0 q^n = a_0 \frac{q^{n+1} - 1}{q - 1} \quad \text{for} \quad q \neq 1, \tag{1.54b}$$

$$s_n = (n+1)a_0 \quad \text{for} \quad q = 1. \tag{1.54c}$$

For $n \to \infty$ (see 7.2.1.1, **2.**, p. 459), there is an *infinite geometric series*, which has a limit if $|q| < 1$, and this limit is called sum s:

$$s = \frac{a_0}{1-q}\,. \tag{1.54d}$$

1.2.4 Special Finite Series

$$1 + 2 + 3 + \cdots + (n-1) + n = \frac{n(n+1)}{2}\,, \tag{1.55}$$

$$p + (p+1) + (p+2) + \cdots + (p+n) = \frac{(n+1)(2p+n)}{2}\,, \tag{1.56}$$

$$1 + 3 + 5 + \cdots + (2n-3) + (2n-1) = n^2\,, \tag{1.57}$$

$$2 + 4 + 6 + \cdots + (2n-2) + 2n = n(n+1)\,, \tag{1.58}$$

$$1^2 + 2^2 + 3^2 + \cdots + (n-1)^2 + n^2 = \frac{n(n+1)(2n+1)}{6}\,, \tag{1.59}$$

$$1^3 + 2^3 + 3^3 + \cdots + (n-1)^3 + n^3 = \frac{n^2(n+1)^2}{4}\,, \tag{1.60}$$

$$1^2 + 3^2 + 5^2 + \cdots + (2n-1)^2 = \frac{n(4n^2-1)}{3}\,, \tag{1.61}$$

$$1^3 + 3^3 + 5^3 + \cdots + (2n-1)^3 = n^2(2n^2-1)\,, \tag{1.62}$$

$$1^4 + 2^4 + 3^4 + \cdots + n^4 = \frac{n(n+1)(2n+1)(3n^2+3n-1)}{30}\,, \tag{1.63}$$

$$1 + 2x + 3x^2 + \cdots + nx^{n-1} = \frac{1 - (n+1)x^n + nx^{n+1}}{(1-x)^2} \quad (x \neq 1)\,. \tag{1.64}$$

1.2.5 Mean Values

(See also 16.3.4.1, **1.**, p. 839 and 16.4, p. 848)

1.2.5.1 Arithmetic Mean or Arithmetic Average

The arithmetic mean of the n quantities $a_1, a_2, \ldots, a_n$ is the expression

$$x_A = \frac{a_1 + a_2 + \cdots + a_n}{n} = \frac{1}{n}\sum_{k=1}^{n} a_k\,. \tag{1.65a}$$

For two values a and b holds:

$$x_A = \frac{a+b}{2}\,. \tag{1.65b}$$

The values a, x_A and b form an arithmetic sequence.

1.2.5.2 Geometric Mean or Geometric Average

The geometric mean of n positive quantities $a_1, a_2, \ldots, a_n$ is the expression

$$x_G = \sqrt[n]{a_1 a_2 \ldots a_n} = \left(\prod_{k=1}^{n} a_k\right)^{\frac{1}{n}}. \tag{1.66a}$$

For two positive values a and b holds

$$x_G = \sqrt{ab}\,. \tag{1.66b}$$

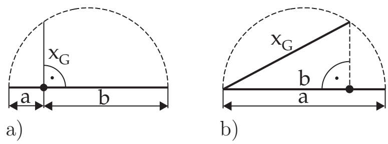


Figure 1.3

The values a, x_G and b form a geometric sequence. If a and b are given line segments, then a segment with length $x_G = \sqrt{ab}$ can be given by the help of one of the constructions shown in **Fig. 1.3a** or in **Fig. 1.3b**.
A special case of the geometric mean is given by dividing a line segment according to the *golden section* (see 3.5.2.3, **3.**, p. 194).

1.2.5.3 Harmonic Mean

The harmonic mean of n quantities $a_1, a_2, \ldots, a_n$ $(a_i \neq 0; i = 1, 2, \ldots, n)$ is the expression

$$x_H = \left[\frac{1}{n}\left(\frac{1}{a_1} + \frac{1}{a_2} + \cdots + \frac{1}{a_n}\right)\right]^{-1} = \left[\frac{1}{n}\sum_{k=1}^{n}\frac{1}{a_k}\right]^{-1}. \tag{1.67a}$$

For two values a and b holds

$$x_H = \left[\frac{1}{2}\left(\frac{1}{a} + \frac{1}{b}\right)\right]^{-1}, \quad x_H = \frac{2ab}{a+b}. \tag{1.67b}$$

1.2.5.4 Quadratic Mean

The *quadratic mean* of n quantities $a_1, a_2, \ldots, a_n$ is the expression

$$x_Q = \sqrt{\frac{1}{n}({a_1}^2 + {a_2}^2 + \cdots + {a_n}^2)} = \sqrt{\frac{1}{n}\sum_{k=1}^{n} a_k^2}\,. \tag{1.68a}$$

For two values a and b holds

$$x_Q = \sqrt{\frac{a^2 + b^2}{2}}\,. \tag{1.68b}$$

The quadratic mean is important in the theory of observational error (see 16.4, p. 848).

1.2.5.5 Relations Between the Means of Two Positive Values

For $x_A = \dfrac{a+b}{2}$, $x_G = \sqrt{ab}$, $x_H = \dfrac{2ab}{a+b}$, $x_Q = \sqrt{\dfrac{a^2+b^2}{2}}$ we have

1. if $\boldsymbol{a < b}$, then

$$a < x_H < x_G < x_A < x_Q < b\,, \tag{1.69a}$$

2. if $\boldsymbol{a = b}$, then

$$a = x_A = x_G = x_H = x_Q = b\,. \tag{1.69b}$$

1.3 Business Mathematics

Business calculations are based on the use of arithmetic and geometric series, on formulas (1.52a)–(1.52c) and (1.54a)–(1.54d). However these applications in banking are so varied and special that a special discipline has developed using specific terminology. So business arithmetic is not confined only to the calculation of the principal by compound interest or the calculation of annuities. It also includes the calculation of interest, repayments, amortization, calculation of instalment payments, annuities, depreciation, effective interest yield and the yield on investment. Basic concepts and formulas for calculations are discussed below. For studying financial mathematics in detail, you will have to consult the relevant literature on the subject (see [1.2], [1.8]).

Insurance mathematics and *risk theory* use the methods of probability theory and mathematical statistics, and they represent a separate discipline, so they don't be discussed here (see [1.4], [1.5]).

1.3.1 Calculation of Interest or Percentage

1.3.1.1 Percentage or Interest

The expression *p percent of K* means $\frac{p}{100}K$, where K denotes the *principal* in business mathematics. The symbol for percent is %, i.e., the following equalities hold:

$$p\% = \frac{p}{100} \qquad \text{or} \qquad 1\% = 0.01. \tag{1.70}$$

1.3.1.2 Increment

If K is raised by $p\%$, the increased value is

$$\tilde{K} = K\left(1 + \frac{p}{100}\right). \tag{1.71}$$

Relating the *increment* $K\frac{p}{100}$ to the new value $\tilde{K}$, the proportion is $K\frac{p}{100} : \tilde{K} = \tilde{p} : 100$, so $\tilde{K}$ contains

$$\tilde{p} = \frac{p \cdot 100}{100 + p} \tag{1.72}$$

percent of increment.

■ If an article has a value of € 200 and a 15% extra charge is added, the final value is € 230. This price contains $\tilde{p} = \frac{15 \cdot 100}{115} = 13.04$ percent increment for the user.

1.3.1.3 Discount or Reduction

Reducing the value K by $p\%$ *rebate* yields the reduced value

$$\tilde{K} = K\left(1 - \frac{p}{100}\right). \tag{1.73}$$

Comparing the reduction $K\frac{p}{100}$ to the new value $\tilde{K}$ gives

$$\tilde{p} = \frac{p \cdot 100}{100 - p} \tag{1.74}$$

percent of rebate.

■ If an article has a value € 300, and they give a 10% discount, it will be sold for € 270. This price contains $\tilde{p} = \frac{10 \cdot 100}{90} = 11.11$ percent rebate for the buyer.

1.3.2 Calculation of Compound Interest

1.3.2.1 Interest

Interest is either payment for the use of a loan or it is a revenue realized from a receivable. For a principal K, placed for a whole *period of interest* (usually one year),

$$K\frac{p}{100} \tag{1.75}$$

interest is paid at the end of the period of interest. Here p is the *rate of interest for the period of interest*, and one says that $p\%$ interest is paid for the principal K.

1.3.2.2 Compound Interest

Compound interest is computed on the principal and on any interest earned that has not been paid or withdrawn. It is the return on the principal for two or more time periods. The interest of the principal increased by interest is called *compound interest*.
In the following different cases are discussed which depend on how the principal is changing.

1. Single Deposit

Compounded annually the principal K increases after n years up to the final value K_n. At the end of the n-th year this value is:

$$K_n = K\left(1+\frac{p}{100}\right)^n. \tag{1.76}$$

For a briefer notation the substitution $1+\frac{p}{100}=q$ is used and q is called the *accumulation factor* or *growth factor*.
Interest may be compounded for any period of time: annually, half-annually, monthly, daily, and so on. Dividing the year into m equal interest periods the interest will be added to the principal K at the end of every period. Then the interest is $K\frac{p}{100m}$ for one interest period, and the principal increases after n years with m interest periods up to the value

$$K_{m\cdot n} = K\left(1+\frac{p}{100m}\right)^{m\cdot n}. \tag{1.77}$$

The quantity $\left(1+\frac{p}{100}\right)$ is known as the *nominal rate*, and $\left(1+\frac{p}{100m}\right)^m$ as the *effective rate*.

■ A principal of € 5000, with a nominal interest 7.2% annually, increases within 6 years **a)** compounded annually to $K_6 = 5000(1+0.072)^6 =$ € 7588.20, **b)** compounded monthly to $K_{72} = 5000(1+0.072/12)^{72} =$ € 7691.74.

2. Regular Deposits

Suppose depositing the same amount E in equal intervals. Such an interval must be equal to an interest period. The depositions can be made at the beginning of the interval, or at the end of the interval. At the end of the n-th interest period the balance K_n is

a) Depositing at the Beginning:

$$K_n = Eq\frac{q^n-1}{q-1}. \tag{1.78a}$$

b) Depositing at the End:

$$K_n = E\frac{q^n-1}{q-1}. \tag{1.78b}$$

3. Depositing in the Course of the Year

A year or an interest period is divided into m equal parts. At the beginning or at the end of each of these time periods the same amount E is deposited and bears interest until the end of the year. In this way, after one year the balance K_1 is

a) Depositing at the Beginning:

$$K_1 = E\left[m + \frac{(m+1)p}{200}\right]. \quad (1.79a)$$

b) Depositing at the End:

$$K_1 = E\left[m + \frac{(m-1)p}{200}\right]. \quad (1.79b)$$

In the second year the total K_1 bears interest, and further deposits and interests are added like in the first year, so after n years the balance K_n for midterm deposits and yearly interest payment is:

a) Depositing at the Beginning:

$$K_n = E\left[m + \frac{(m+1)p}{200}\right]\frac{q^n - 1}{q - 1}. \quad (1.80a)$$

b) Depositing at the End:

$$K_n = E\left[m + \frac{(m-1)p}{200}\right]\frac{q^n - 1}{q - 1}. \quad (1.80b)$$

■ At a yearly rate of interest $p = 5.2\%$ a depositor deposits € 1000 at the end of every month. After how many years will it reach the balance € 500 000?

From (1.80b), for instance, from $500\,000 = 1000\left[12 + \frac{11 \cdot 5.2}{200}\right] \cdot \frac{1.052^n - 1}{0.052}$, follows the answer, $n = 22.42$ years.

1.3.3 Amortization Calculus

1.3.3.1 Amortization

Amortization is the repayment of credits. The assumptions:

1. For a *debt* S the debtor is charged at $p\%$ interest at the end of an interest period.

2. After N interest period the debt is completely repaid.

The charge of the debtor consists of interest and principal repayment for every interest period. If the interest period is one year, the amount to be paid during the whole year is called an *annuity*.

There are different possibilities for a debtor. For instance, the repayments can be made at the interest date, or meanwhile; the amount of repayment can be different time by time, or it can be constant during the whole term.

1.3.3.2 Equal Principal Repayments

The amortization instalments are paid during the year, but no midterm compound interest is calculated. The following notation should be used:

- S debt (interest payment at the end of a period with $p\%$),
- $T = \frac{S}{mN}$ principal repayment ($T = \text{const}$),
- m number of repayments during one interest period,
- N number of interest periods until the debt is fully repaid.

Besides the principal repayments the debtor also has to pay the interest charges:

a) Interest Z_n for the n-th Interest Period:

$$Z_n = \frac{pS}{100}\left[1 - \frac{1}{N}\left(n - \frac{m+1}{2m}\right)\right]. \quad (1.81a)$$

b) Total Interest Z to be Paid for a Debt S, mN Times, During N Interest Periods with an Interest Rate $p\%$:

$$Z = \sum_{n=1}^{N} Z_n = \frac{pS}{100}\left[\frac{N-1}{2} + \frac{m+1}{2m}\right]. \quad (1.81b)$$

■ A debt of € 60 000 has a yearly interest rate of 8%. The principal repayment of € 1000 for 60 months should be paid at the end of the months. How much is the actual interest at the end of each year? The interest for every year is calculated by (1.81a) with $S = 60000$, $p = 8$, $N = 5$ and $m = 12$. They are enumerated in the annexed table.

1. year:	$Z_1 =$	€ 4360
2. year:	$Z_2 =$	€ 3400
3. year:	$Z_3 =$	€ 2440
4. year:	$Z_4 =$	€ 1480
5. year:	$Z_5 =$	€ 520
	$Z =$	€ 12200

The total interest can be calculated also by (1.81b) as $Z = \frac{8 \cdot 60000}{100}\left[\frac{5-1}{2}+\frac{13}{24}\right] = € \, 12\,200$.

1.3.3.3 Equal Annuities

For equal principal repayments $T = \frac{S}{mN}$ the interest payable decreases over the course of time (see the previous example). In contrast to this, in the case of equal *annuities* the same amount is repaid for every interest period. A constant annuity A containing the principal repayment and the interest is repaid, i.e., the charge of the debtor is constant during the whole period of repayment.
With the notation
- S debt (interest payment of $p\%$ at the end of a period),
- A annuity for every interest period (A const),
- a one instalment paid m times per interest period (a const),
- $q = 1 + \frac{p}{100}$ the accumulation factor,

after n interest periods the remaining outstanding debt S_n is:

$$S_n = S\,q^n - a\left[m + \frac{(m-1)p}{200}\right]\frac{q^n-1}{q-1}\,. \tag{1.82}$$

Here the term Sq^n denotes the value of the debt S after n interest periods with compound interest (see (1.76)). The second term in (1.82) gives the value of the midterm repayments a with compound interest (see (1.80b) with $E = a$). For the annuity holds

$$A = a\left[m + \frac{(m-1)p}{200}\right]. \tag{1.83}$$

Here paying A once means the same as paying a m times. From (1.83) it follows that $A \geq ma$. Because after N interest periods the debt must be completely repaid, from (1.82) for $S_N = 0$ considering (1.83) for the annuity holds:

$$A = S\,q^N\frac{q-1}{q^N-1} = S\frac{q-1}{1-q^{-N}}\,. \tag{1.84}$$

To solve a problem of business mathematics, from (1.84), any of the quantities A, S, q or N can be expressed, if the others are known.

■ **A:** A loan of € 60 000 bears 8% interest per year, and is to be repaid over 5 years in equal instalments. How much is the yearly annuity A and the monthly instalment a? From (1.84) and (1.83) we get: $A = 60\,000\frac{0.08}{1-\frac{1}{1.08^5}} = € \, 15\,027.39$, $a = \frac{15027.39}{12+\frac{11\cdot 8}{200}} = € \, 1207.99$.

■ **B:** A loan of $S =$ € 100 000 is to be repaid during $N = 8$ years in equal annuities with an interest rate of 7.5%. At the end of every year € 5000 extra repayment must be made. How much will the monthly instalment be? For the annuity A per year according to (1.84) follows $A = 100\,000\frac{0.075}{1-\frac{1}{1.075^8}} =$ € 17 072.70. Because A consists of 12 monthly instalments a, and because of the € 5000 extra payment at the end of the year, from (1.83) $A = a\left[12 + \frac{11\cdot 7.5}{200}\right] + 5000 = 17\,072.70$ holds, so the monthly charge is $a =$ € 972.62.

1.3.4 Annuity Calculations

1.3.4.1 Annuities

If a series of payments is made regularly at the same time intervals, in equal or varying amounts, at the beginning or at the end of the interval, it is called *annuity payments*. To distinguish are:

a) Payments on an Account The periodic payments, called *rents*, are paid on an account and bear compound interest. Therefore the formulas of 1.3.2 are to be used.

b) Receipt of Payments The payments of rent are made from capital bearing compound interest. Here the formulas of the annuity calculations in 1.3.3 are to be used, where the annuities are called rents. If no more than the actual interest is paid as a rent, it is called a *perpetual annuity*.

Rent payments (deposits and payoffs) can be made at the interest terms, or at shorter intervals during the period of interest, i.e. in the course of the year.

1.3.4.2 Future Amount of an Ordinary Annuity

The date of the interest calculations and the payments should coincide. The interest is calculated at $p\%$ compound interest, and the payments (rents) on the account are always the same, R. The *future value of the ordinary annuity* R_n, i.e., the amount to which the regular deposits increase after n periods amounts to:

$$R_n = R\frac{q^n - 1}{q - 1} \quad \text{with} \quad q = 1 + \frac{p}{100}\,. \tag{1.85}$$

The *present value of an ordinary annuity* R_0 is the amount which should be paid at the beginning of the first interest period (one time) to reach the final value R_n with compound interest during n periods:

$$R_0 = \frac{R_n}{q^n} \qquad \text{with} \quad q = 1 + \frac{p}{100}\,. \tag{1.86}$$

■ A man claims € 5000 at the end of every year for 10 years from a firm. Before the first payment the firm declares bankruptcy. Only the present value of the ordinary annuity R_0 can be asked from the administration of the bankrupt's estate. With an interest of 4% per year the man gets:

$$R_0 = \frac{1}{q^n}R\frac{q^n - 1}{q - 1} = R\frac{1 - q^{-n}}{q - 1} = 5000\frac{1 - 1.04^{-10}}{0.04} = € \, 40\,554.48.$$

1.3.4.3 Balance after n Annuity Payments

For ordinary annuity payments capital K is at our disposal bearing $p\%$ interest. After every interest period an amount r is paid. The balance K_n after n interest periods, i.e., after n rent payments, is:

$$K_n = Kq^n - R_n = Kq^n - r\frac{q^n - 1}{q - 1} \quad \text{with} \quad q = 1 + \frac{p}{100}\,. \tag{1.87a}$$

Conclusions from (1.87a):

$$r = K\frac{p}{100} \tag{1.87b}$$

Consequently $K_n = K$ holds, so the capital does not change. This is the case of *perpetual annuity*.

$$r > K\frac{p}{100} \tag{1.87c}$$

The capital will be completely used up after N rent payments. From (1.87a) it follows for $K_N = 0$:

$$K = \frac{r}{q^N}\frac{q^N - 1}{q - 1}\,. \tag{1.87d}$$

If midterm interest is calculated and midterm rents are paid, and the original interest period is divided into m equal intervals, then in the formulas (1.85)–(1.87a) n is replaced by mn and accordingly $q = 1 + \frac{p}{100}$ by $q = 1 + \frac{p}{100m}$.

■ What amount must be deposited monthly at the end of the month for 20 years, from which a rent of € 2000 should be paid monthly for 20 years, and the interest period is one month with an interest rate of 0.5%.
From (1.87d) follows for $n = 20 \cdot 12 = 240$ the sum K which is necessary for the required payments: $K = \frac{2000}{1.005^{240}} \frac{1.005^{240} - 1}{0.005} =$ € 279 161.54. The necessary monthly deposits R are given by (1.85): $R_{240} = 279\,161.54 = R\frac{1.005^{240} - 1}{0.005}$, i.e., $R =$ € 604.19.

1.3.5 Depreciation

1.3.5.1 Methods of Depreciation

Depreciation is the term most often used to indicate that assets have declined in service potential in a given year either due to obsolescence or physical factors. Depreciation is a method whereby the *original (cost) value* at the beginning of the reporting year is reduced to the *residual value* at year-end. The following concepts are used:
- A depreciation base,
- N useful life (given in years),
- R_n residual value after n years ($n \le N$),
- a_n $(n = 1, 2, \ldots, N)$ depreciation rate in the n-th year.

The methods of depreciation differ from each other depending on the *amortization rate*:
- *straight-line method*, i.e., equal yearly rates,
- *decreasing-charge method*, i.e., decreasing yearly rates.

1.3.5.2 Straight-Line Method

The yearly depreciations are constant, i.e., for amortization rates a_n and the remaining value R_n after n years follows:

$$a_n = \frac{A - R_N}{N} = a, \quad (1.88) \qquad R_n = A - n\frac{A - R_N}{N} \quad (n = 1, 2, \ldots, N). \quad (1.89)$$

Substituting $R_N = 0$, then the value of the given thing is reduced to zero after N years, i.e., it is totally depreciated.

■ The purchase price of a machine is $A =$ € 50 000. In 5 years it should be depreciated to a value $R_5 =$ € 10 000.
Linear depreciation according to (1.88) and (1.89) yields the annexed *amortization schedule*:
It shows that the percentage of accumulated depreciation with respect to the actual initial value is increasing.

Year	Depreciation base	Depreciation expense	Residual value	Cumulated depr. in % of the depr. base
1	50 000	8000	42 000	16.0
2	42 000	8000	34 000	19.0
3	34 000	8000	26 000	23.5
4	26 000	8000	18 000	30.8
5	18 000	8000	10 000	44.4

1.3.5.3 Arithmetically Declining Balance Depreciation

In this case the depreciation is not constant. It is decreasing yearly by the same amount d, by the so-called *multiple*. For depreciation in the n-th year follows:

$$a_n = a_1 - (n-1)d \quad (n = 2, 3, \ldots, N+1;\ a_1 \text{ and } d \text{ are given}). \quad (1.90)$$

Considering the equality $A - R_N = \sum\limits_{n=1}^{N} a_n$ from the previous equation it follows that:

$$d = \frac{2[Na_1 - (A - R_N)]}{N(N-1)}. \quad (1.91)$$

For $d = 0$ follows the special case of straight-line depreciation. If $d > 0$, it follows from (1.91) that

$$a_1 > \frac{A - R_N}{N} = a, \tag{1.92}$$

where a is the depreciation rate for straight-line depreciation. The first depreciation rate a_1 of the arithmetically-declining balance depreciation must satisfy the following inequality:

$$\frac{A - R_N}{N} < a_1 < 2\frac{A - R_N}{N}\,. \tag{1.93}$$

■ A machine of € 50 000 purchase price is to be depreciated to the value of € 10 000 within 5 years by arithmetically declining depreciation. In the first year € 15 000 should be depreciated.

The annexed depreciation schedule is calculated by the given formulas, and it shows that with the exception of the last rate the percentage of depreciation is fairly equal.

Year	Depretiation base	Depreciation expense	Residual value	Depreciation in % of depr. base
1	50 000	15 000	35 000	30.0
2	35 000	11 500	23 500	32.9
3	23 500	8 000	15 500	34.0
4	15 500	4 500	11 000	29.0
5	11 000	1 000	10 000	9.1

1.3.5.4 Digital Declining Balance Depreciation

Digital depreciation is a special case of arithmetically declining depreciation. Here it is required that the last depreciation rate a_N should be equal to the multiple d. From $a_N = d$ it follows that

$$d = \frac{2(A - R_N)}{N(N+1)}\,, \tag{1.94a}$$

$$a_1 = Nd,\ a_2 = (N-1)d,\ \ldots, a_N = d. \tag{1.94b}$$

■ The purchase price of a machine is € $A = 50\,000$. This machine is to be depreciated in 5 years to the value $R_5 =$ € 10 000 by digital depreciation.

The annexed depreciation schedule, calculated by the given formulas, shows that the percentage of the depreciation is fairly equal.

Year	Depreciation base	Depreciation expense	Residual value	Depreciation in % of the depr. base
1	50 000	$a_1 = 5d = 13\,335$	36 665	26.7
2	36 665	$a_2 = 4d = 10\,668$	25 997	29.1
3	25 997	$a_3 = 3d = 8\,001$	17 996	30.8
4	17 996	$a_4 = 2d = 5\,334$	12 662	29.6
5	12 662	$a_5 = d = 2\,667$	9 995	21.1

1.3.5.5 Geometrically Declining Balance Depreciation

Consider geometrically declining depreciation where p% of the actual value is depreciated every year. For the residual value R_n after n years holds:

$$R_n = A\left(1 - \frac{p}{100}\right)^n \quad (n = 1, 2, \ldots)\,. \tag{1.95}$$

Usually A (the acquisition cost) is given. The useful life of the asset is N years long. If from the quantities R_N, p and N, two is given, the third one can be calculated by the formula (1.95).

■ **A:** A machine with a purchase value € 50 000 is to be geometrically depreciated yearly by 10%. After how many years will its value drop below € 10 000 for the first time? Based on (1.95), yields

$$N = \frac{\ln(10\,000/50\,000)}{\ln(1 - 0.1)} = 15.27 \text{ years.}$$

■ **B:** For a purchase price of $A =$ € 1000 the residual value R_n should be represented for $n = 1, 2, \ldots, 10$ years by a) straight-line, b) arithmetically declining, c) geometrically declining depreciation. The results are shown in **Fig. 1.4**.

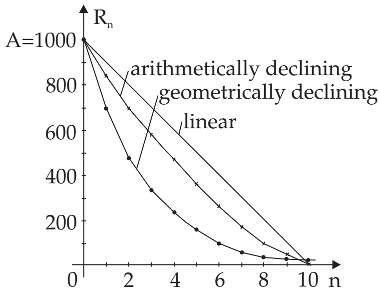


Figure 1.4

1.3.5.6 Depreciation with Different Types of Depreciation Account

Since in the case of geometrically declining depreciation the residual value cannot become equal to zero for a finite n, it is reasonable after a certain time, e.g., after m years, to switch over to straight-line depreciation. m is to be determined to an amount that from this time on the geometrically declining depreciation rate is smaller than the straight-line depreciation rate. From this requirement it follows that:

$$m > N - \frac{100}{p}\,. \tag{1.96}$$

Here m is the last year of geometrically declining depreciation and N is the last year of linear depreciation when the residual value becomes zero.

■ A machine with a purchase value of € 50 000 is to be depreciated to zero within 15 years, for m years by geometrically declining depreciation with 14% of the residual value, then with the straight-line method. From (1.96) follows $m > 15 - \frac{100}{14} = 7.76$, i.e., after $m = 8$ years it is reasonable to switch over to straight-line depreciation.

1.4 Inequalities

1.4.1 Pure Inequalities

1.4.1.1 Definitions

1. Inequalities

Inequalities are comparisons of two real algebraic expressions represented by one of the following signs:

Type I	$>$	("greater")	Type II	$<$	("smaller")
Type III	$\neq$	("not equal")	Type IIIa	$<>$	("greater or smaller")
Type IV	$\geq$	("greater or equal")	Type IVa	$\not<$	("not smaller")
Type V	$\leq$	("smaller or equal")	Type Va	$\not>$	("not greater")

The notation III and IIIa, IV and IVa, and V and Va have the same meaning, so they can be replaced by each other. The notation III can also be used for those types of quantities for which the notions of "greater" or "smaller" cannot be defined, for instance for complex numbers or vectors, but in this case it cannot be replaced by IIIa.

2. Identical Inequalities, Inequalities of the Same and of the Opposite Sense, Equivalent Inequalities

1. Identical Inequalities are valid for arbitrary values of the letters contained in them.

2. Inequalities of the Same Sense belong to the same type from the first two, i.e., both belong to type I or both belong to type II.

3. Inequalities of the Opposite Sense belong to different types of the first two, i.e., one to type I, the other to type II.

4. Equivalent Inequalities are inequalities if they are valid exactly for the same values of the unknowns contained in them.

3. Solution of Inequalities

Similarly to equalities, inequalities can contain unknown quantities which are usually denoted by the last letters of the alphabet. The *solution of an inequality* or a system of inequalities means the determination of the limits for the unknowns between which they can change, keeping the inequality or system

of inequalities true.

Solutions can be looked for any kind of inequality; mostly *pure inequalities* of type I and II are to be solved.

1.4.1.2 Properties of Inequalities of Type I and II

1. Change the Sense of the Inequality

$$\text{If} \quad a > b \text{ holds,} \quad \text{then} \quad b < a \quad \text{is valid,} \tag{1.97a}$$

$$\text{if} \quad a < b \text{ holds,} \quad \text{then} \quad b > a \quad \text{is valid.} \tag{1.97b}$$

2. Transitivity

$$\text{If} \quad a > b \quad \text{and} \quad b > c \text{ hold,} \quad \text{then} \quad a > c \quad \text{is valid;} \tag{1.98a}$$

$$\text{if} \quad a < b \quad \text{and} \quad b < c \text{ hold,} \quad \text{then} \quad a < c \quad \text{is valid.} \tag{1.98b}$$

3. Addition and Subtraction of a Quantity

$$\text{If} \quad a > b \text{ holds,} \quad \text{then} \quad a \pm c > b \pm c \quad \text{is valid;} \tag{1.99a}$$

$$\text{if} \quad a < b \text{ holds,} \quad \text{then} \quad a \pm c < b \pm c \quad \text{is valid.} \tag{1.99b}$$

By adding or subtracting the same amount to the both sides of inequality, the sense of the inequality does not change.

4. Addition of Inequalities

$$\text{If} \quad a > b \quad \text{and} \quad c > d \text{ hold,} \quad \text{then} \quad a + c > b + d \quad \text{is valid;} \tag{1.100a}$$

$$\text{if} \quad a < b \quad \text{and} \quad c < d \text{ hold,} \quad \text{then} \quad a + c < b + d \quad \text{is valid.} \tag{1.100b}$$

Two inequalities of the same sense can be added.

5. Subtraction of Inequalities

$$\text{If} \quad a > b \quad \text{and} \quad c < d \text{ hold,} \quad \text{then} \quad a - c > b - d \quad \text{is valid;} \tag{1.101a}$$

$$\text{if} \quad a < b \quad \text{and} \quad c > d \text{ hold,} \quad \text{then} \quad a - c < b - d \quad \text{is valid.} \tag{1.101b}$$

Inequalities of the opposite sense can be subtracted; the result keeps the sense of the first inequality. Subtracting inequalities of the same sense is not allowed.

6. Multiplication and Division of an Inequality by a Quantity

$$\text{If} \quad a > b \quad \text{and} \quad c > 0 \text{ hold,} \quad \text{then} \quad ac > bc \quad \text{and} \quad \frac{a}{c} > \frac{b}{c} \text{ are valid,} \tag{1.102a}$$

$$\text{if} \quad a < b \quad \text{and} \quad c > 0 \text{ hold,} \quad \text{then} \quad ac < bc \quad \text{and} \quad \frac{a}{c} < \frac{b}{c} \text{ are valid,} \tag{1.102b}$$

$$\text{if} \quad a > b \quad \text{and} \quad c < 0 \text{ hold,} \quad \text{then} \quad ac < bc \quad \text{and} \quad \frac{a}{c} < \frac{b}{c} \text{ are valid,} \tag{1.102c}$$

$$\text{if} \quad a < b \quad \text{and} \quad c < 0 \text{ hold,} \quad \text{then} \quad ac > bc \quad \text{and} \quad \frac{a}{c} > \frac{b}{c} \text{ are valid.} \tag{1.102d}$$

Multiplication or division of both sides of an inequality by a positive value does not change the sense of the inequality. Multiplication or division by a negative value changes the sense of the inequality.

7. Inequalities and Reciprocal Values

$$\text{If} \quad 0 < a < b \quad \text{or} \quad a < b < 0 \text{ hold,} \quad \text{then} \quad \frac{1}{a} > \frac{1}{b} \text{ is valid.} \tag{1.103}$$

1.4.2 Special Inequalities

1.4.2.1 Triangle Inequality for Real Numbers

For arbitrary real numbers $a, b, a_1, a_2, \ldots, a_n$, there are the inequalities

$$|a+b| \le |a| + |b|\,; \qquad |a_1 + a_2 + \cdots + a_n| \le |a_1| + |a_2| + \cdots + |a_n|\,. \tag{1.104}$$

The absolute value of the sum of two or more real numbers is less than or equal to the sum of their absolute values. The equality holds only if the summands have the same sign.

1.4.2.2 Triangle Inequality for Complex Numbers

For n complex numbers $z_1, z_2, \ldots, z_n \in \mathbb{C}$

$$\left|\sum_{k=1}^{n} z_k\right| = |z_1 + z_2 + \cdots + z_n| \le |z_1| + |z_2| + \cdots + |z_n| = \sum_{k=1}^{n} |z_k|. \tag{1.105}$$

1.4.2.3 Inequalities for Absolute Values of Differences of Real and Complex Numbers

For arbitrary real numbers $a, b \in \mathbb{R}$, there are the inequalities

$$||a| - |b|| \le |a - b| \le |a| + |b|. \tag{1.106}$$

The absolute value of the difference of two real numbers is less than or equal to the sum of their absolute values, but greater than or equal to the absolute value of the difference of their absolute values. For two arbitrary complex numbers $z_1, z_2 \in \mathbb{C}$

$$||z_1| - |z_2|| \le |z_1 - z_2| \le |z_1| + |z_2|\,. \tag{1.107}$$

1.4.2.4 Inequality for Arithmetic and Geometric Means

$$\frac{a_1 + a_2 + \cdots + a_n}{n} \ge \sqrt[n]{a_1 a_2 \cdots a_n} \quad \text{for} \quad a_i > 0\,. \tag{1.108}$$

The arithmetic mean of n positive numbers is greater than or equal to their geometric mean. Equality holds only if all the n numbers are equal.

1.4.2.5 Inequality for Arithmetic and Quadratic Means

$$\left|\frac{a_1 + a_2 + \cdots + a_n}{n}\right| \le \sqrt{\frac{{a_1}^2 + {a_2}^2 + \cdots + {a_n}^2}{n}}\,. \tag{1.109}$$

The absolute value of the arithmetic mean of numbers is less than or equal to their quadratic mean.

1.4.2.6 Inequalities for Different Means of Real Numbers

For the harmonic, geometric, arithmetic, and quadratic means of two positive real numbers a and b with $a < b$ the following inequalities hold (see also 1.2.5.5, p. 20):

$$a < x_H < x_G < x_A < x_Q < b. \tag{1.110a}$$

Here

$$x_A = \frac{a+b}{2}\,,\; x_G = \sqrt{ab}\,,\; x_H = \frac{2ab}{a+b}\,,\; x_Q = \sqrt{\frac{a^2+b^2}{2}}\,. \tag{1.110b}$$

1.4.2.7 Bernoulli's Inequality

For every real number $a \ge -1$ and integer $n \ge 1$ holds

$$(1+a)^n \ge 1 + n\,a\,. \tag{1.111}$$

The equality holds only for $n = 1$, or $a = 0$.

1.4.2.8 Binomial Inequality

For arbitrary real numbers $a, b \in \mathbb{R}$ holds

$$|a\,b| \leq \frac{1}{2}(a^2 + b^2)\,. \tag{1.112}$$

1.4.2.9 Cauchy-Schwarz Inequality

1. Cauchy-Schwarz Inequality for Real Numbers

The Cauchy-Schwarz inequality holds for arbitrary real numbers $a_i, b_j \in \mathbb{R}$:

$$|a_1b_1 + a_2b_2 + \cdots + a_nb_n| \leq \sqrt{{a_1}^2 + {a_2}^2 + \cdots + {a_n}^2}\sqrt{{b_1}^2 + {b_2}^2 + \cdots + {b_n}^2} \tag{1.113a}$$

or

$$(a_1b_1 + a_2b_2 + \cdots + a_nb_n)^2 \leq ({a_1}^2 + {a_2}^2 + \cdots + {a_n}^2)({b_1}^2 + {b_2}^2 + \cdots + {b_n}^2). \tag{1.113b}$$

For two finite sequences of n real numbers, the sum of the pairwise products is less than or equal to the product of the square roots of the sums of the squares of these numbers. Equality holds only if $a_1 : b_1 = a_2 : b_2 = \cdots = a_n : b_n$.

If $n = 3$ and $\{a_1, a_2, a_3\}$ and $\{b_1, b_2, b_3\}$ are considered as vectors in a Cartesian coordinate system, then the Cauchy-Schwarz inequality means that the absolute value of the scalar product of two vectors is less than or equal to the product of absolute values of these vectors. If $n > 3$, then this statement can be extended for vectors in n-dimensional Euclidean space.

2. Cauchy-Schwarz Inequality for Complex Numbers

Considering that for complex numbers $|z|^2 = z^*z$ (z^* is the complex conjugate of z), the inequality (1.113b) is valid also for arbitrary complex numbers $z_i, w_j \in \mathbb{C}$:

$(z_1w_1 + z_2w_2 + \cdots + z_nw_n)^*(z_1w_1 + z_2w_2 + \cdots + z_nw_n)$
$\leq ({z_1}^*z_1 + {z_2}^*z_2 + \cdots + {z_n}^*z_n)({w_1}^*w_1 + {w_2}^*w_2 + \cdots + {w_n}^*w_n)$.

3. Cauchy-Schwarz Inequality for Convergent Infinite Series and Integrals

An analogous statement to (1.113b) is the Cauchy-Schwarz inequality for convergent infinite series and for certain integrals:

$$\left(\sum_{n=1}^{\infty} a_nb_n\right)^2 \leq \left(\sum_{n=1}^{\infty} {a_n}^2\right)\left(\sum_{n=1}^{\infty} {b_n}^2\right), \tag{1.114}$$

$$\left[\int_a^b f(x)\,\varphi(x)\,dx\right]^2 \leq \left(\int_a^b [f(x)]^2\,dx\right)\left(\int_a^b [\varphi(x)]^2\,dx\right). \tag{1.115}$$

1.4.2.10 Chebyshev Inequality

If $a_1, a_2, \ldots, a_n, b_1, b_2, \ldots, b_n$ are real positive numbers, then the following inequalities hold:

$$\left(\frac{a_1 + a_2 + \cdots + a_n}{n}\right)\left(\frac{b_1 + b_2 + \cdots + b_n}{n}\right) \leq \frac{a_1b_1 + a_2b_2 + \cdots + a_nb_n}{n} \tag{1.116a}$$

for $a_1 \leq a_2 \leq \ldots \leq a_n$ and $b_1 \leq b_2 \leq \ldots \leq b_n$,
or $a_1 \geq a_2 \geq \ldots \geq a_n$ and $b_1 \geq b_2 \geq \ldots \geq b_n$,

and

$$\left(\frac{a_1 + a_2 + \cdots + a_n}{n}\right)\left(\frac{b_1 + b_2 + \cdots + b_n}{n}\right) \geq \frac{a_1b_1 + a_2b_2 + \cdots + a_nb_n}{n} \tag{1.116b}$$

for $a_1 \leq a_2 \leq \ldots \leq a_n$ and $b_1 \geq b_2 \geq \ldots \geq b_n$.

For two finite sequences with n positive numbers, the product of the arithmetic means of these sequences is less than or equal to the aritmetic mean of the pairwise products if both sequences are increasing or

both are decreasing; but the inequality is valid in the opposite sense if one of the sequences is increasing and the other one is decreasing.

1.4.2.11 Generalized Chebyshev Inequality

If $a_1, a_2, \ldots, a_n, b_1, b_2, \ldots, b_n$ are real positive numbers, then the following inequalities hold:

$$\sqrt[k]{\frac{{a_1}^k + {a_2}^k + \cdots + {a_n}^k}{n}} \; \sqrt[k]{\frac{{b_1}^k + {b_1}^k + \cdots + {b_n}^k}{n}} \le \sqrt[k]{\frac{(a_1b_1)^k + (a_2b_2)^k + \cdots + (a_nb_n)^k}{n}} \tag{1.117a}$$

for $a_1 \le a_2 \le \ldots \le a_n$ and $b_1 \le b_2 \le \ldots \le b_n$
or $a_1 \ge a_2 \ge \ldots \ge a_n$ and $b_1 \ge b_2 \ge \ldots \ge b_n$

and

$$\sqrt[k]{\frac{{a_1}^k + {a_2}^k + \cdots + {a_n}^k}{n}} \; \sqrt[k]{\frac{{b_1}^k + {b_1}^k + \cdots + {b_n}^k}{n}} \ge \sqrt[k]{\frac{(a_1b_1)^k + (a_2b_2)^k + \cdots + (a_nb_n)^k}{n}} \tag{1.117b}$$

for $a_1 \le a_2 \le \ldots \le a_n$ and $b_1 \ge b_2 \ge \ldots \ge b_n$.

1.4.2.12 Hölder Inequality

1. Hölder Inequality for Series

If p and q are two real numbers such that $\frac{1}{p} + \frac{1}{q} = 1$ is fulfilled, and if $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$ are arbitrary $2n$ complex numbers, then the following inequality holds:

$$\sum_{k=1}^{n} |x_k y_k| \le \left[\sum_{k=1}^{n} |x_k|^p\right]^{\frac{1}{p}} \left[\sum_{k=1}^{n} |y_k|^q\right]^{\frac{1}{q}}. \tag{1.118a}$$

This inequality is also valid for countable infinite pairs of numbers:

$$\sum_{k=1}^{\infty} |x_k y_k| \le \left[\sum_{k=1}^{\infty} |x_k|^p\right]^{\frac{1}{p}} \left[\sum_{k=1}^{\infty} |y_k|^q\right]^{\frac{1}{q}}, \tag{1.118b}$$

where from the convergence of the series on the right-hand side the convergence of the left-hand side follows.

2. Hölder Inequality for Integrals

If $f(x)$ and $g(x)$ are two measurable functions on the measure space $(\mathrm{X}, \mathcal{A}, \mu)$ (see 12.9.2, p. 695), then the following inequality holds:

$$\int_X |f(x)g(x)| d\mu \le \left[\int_X |f(x)|^p \, d\mu\right]^{\frac{1}{p}} \left[\int_X |g(x)|^q \, d\mu\right]^{\frac{1}{q}}. \tag{1.118c}$$

1.4.2.13 Minkowski Inequality

1. Minkowski Inequality for Series

If $p \ge 1$ holds, and $\{x_k\}_{k=1}^{k=\infty}$ and $\{y_k\}_{k=1}^{\infty}$ with $x_k, y_k \in \mathbb{C}$ are two sequences of numbers, then holds:

$$\left[\sum_{k=1}^{\infty} |x_k + y_k|^p\right]^{\frac{1}{p}} \le \left[\sum_{k=1}^{\infty} |x_k|^p\right]^{\frac{1}{p}} + \left[\sum_{k=1}^{\infty} |y_k|^p\right]^{\frac{1}{p}}. \tag{1.119a}$$

2. Minkowski Inequality for Integrals

If $f(x)$ and $g(x)$ are two measurable functions on the measure space $(\mathrm{X}, \mathcal{A}, \mu)$ (see 12.9.2, p. 695), then holds:

$$\left[\int_X |f(x)+g(x)|^p d\mu\right]^{\frac{1}{p}} \le \left[\int_X |f(x)|^p d\mu\right]^{\frac{1}{p}} + \left[\int_X |g(x)|^p d\mu\right]^{\frac{1}{p}}. \qquad (1.119b)$$

1.4.3 Solution of Linear and Quadratic Inequalities

1.4.3.1 General Remarks

During the solution of an inequality it is transformed into equivalent inequalities step by step. Similarly to the solution of an equation the same expression can be added to both sides; formally, it may seem that a summand is brought from one side to the other, changing its sign. Furthermore one can multiply or divide both sides of an inequality by a non-zero expression, where the inequality keeps its sense if this expression has a positive value, and changes its sense if this expression has a negative value. An inequality of first degree can always be transformed into the form

$$ax > b. \qquad (1.120)$$

The simplest form of an inequality of second degree is

$$x^2 > m \quad (1.121a) \qquad \text{or} \quad x^2 < m \qquad (1.121b)$$

and in the general case it has the form

$$ax^2 + bx + c > 0 \quad (1.122a) \qquad \text{or} \quad ax^2 + bx + c < 0. \qquad (1.122b)$$

1.4.3.2 Linear Inequalities

The linear inequality of first degree (1.120) has the solution

$$x > \frac{b}{a} \ \text{for}\ a > 0 \quad (1.123a) \qquad \text{and} \quad x < \frac{b}{a} \ \text{for}\ a < 0. \qquad (1.123b)$$

■ $5x + 3 < 8x + 1, \quad 5x - 8x < 1 - 3, \quad -3x < -2, \quad x > \frac{2}{3}.$

1.4.3.3 Quadratic Inequalities

Inequalities of second degree in the form

$$x^2 > m \quad (1.124a) \qquad \text{and} \quad x^2 < m \qquad (1.124b)$$

have solutions

a) $x^2 > m$: For $m \ge 0$ the solution is $x > \sqrt{m}$ and $x < -\sqrt{m}$ $(|x| > \sqrt{m})$, (1.125a)

for $m < 0$ the inequality obviously holds for any x. (1.125b)

b) $x^2 < m$: For $m > 0$ the solution is $-\sqrt{m} < x < \sqrt{m}$ $(|x| < \sqrt{m})$, (1.126a)

for $m \le 0$ there is no solution. (1.126b)

1.4.3.4 General Case for Inequalities of Second Degree

$$ax^2 + bx + c > 0 \quad (1.127a) \qquad \text{or} \quad ax^2 + bx + c < 0. \qquad (1.127b)$$

First dividing the inequality by a. If $a < 0$ then the sense of the inequality changes, but in any case it will have the form

$$x^2 + px + q < 0 \quad (1.127c) \qquad \text{or} \quad x^2 + px + q > 0. \qquad (1.127d)$$

By completing the square it follows that

$$\left(x+\frac{p}{2}\right)^2 < \left(\frac{p}{2}\right)^2 - q \qquad (1.127e) \qquad \text{or} \qquad \left(x+\frac{p}{2}\right)^2 > \left(\frac{p}{2}\right)^2 - q. \qquad (1.127f)$$

Denoting $x+\frac{p}{2}$ by z and $\left(\frac{p}{2}\right)^2 - q$ by m, the inequalities

$$z^2 < m \qquad (1.128a) \qquad \text{or} \qquad z^2 > m \qquad (1.128b)$$

can be obtained. Solving these inequalities yields the values for x.

■ **A:** $-2x^2+14x-20>0,\ x^2-7x+10<0,\ \left(x-\frac{7}{2}\right)^2<\frac{9}{4},\ -\frac{3}{2}<x-\frac{7}{2}<\frac{3}{2},$
$-\frac{3}{2}+\frac{7}{2}<x<\frac{3}{2}+\frac{7}{2}.$
The solution is $2<x<5$.

■ **B:** $x^2+6x+15>0,\ (x+3)^2>-6.$ The inequality holds identically.

■ **C:** $-2x^2+14x-20<0,\ \left(x-\frac{7}{2}\right)^2>\frac{9}{4},\ x-\frac{7}{2}>\frac{3}{2}$ and $x-\frac{7}{2}<-\frac{3}{2}.$
The solution intervals are $x>5$ and $x<2$.

1.5 Complex Numbers

1.5.1 Imaginary and Complex Numbers

1.5.1.1 Imaginary Unit

The imaginary unit is denoted by i, which represents a number different from any real number, and whose square is equal to -1. In electronics, instead of i the letter j is usually used to avoid accidently confusing it with the intensity of current, also denoted by i. The introduction of the *imaginary unit* leads to the *generalization of the notion of numbers* to the *complex numbers*, which play a very important role in algebra and analysis. The complex numbers have several interpretations in geometry and physics.

1.5.1.2 Complex Numbers

The *algebraic form of a complex number* is

$$z = a + \mathrm{i}\,b. \qquad (1.129a)$$

When a and b take all possible real values, then one gets all possible complex numbers z. The number a is the *real part*, the number b is the *imaginary part* of the number z:

$$a = \mathrm{Re}(z), \quad b = \mathrm{Im}(z). \qquad (1.129b)$$

For $b=0$ it is $z=a$, so the real numbers form a subset of the complex numbers. For $a=0$ it is $z=\mathrm{i}\,b$, which is a "pure imaginary number".
The total set of complex numbers is denoted by $\mathbb{C}$.

Remark: Functions $w=f(z)$ with complex variable $z=x+\mathrm{i}\,y$ will be discussed in function theory (see 14.1, p. 731 ff).

1.5.2 Geometric Representation

1.5.2.1 Vector Representation

Similarly to the representation of the real numbers on the numerical axis, the complex numbers can be represented as points in the so-called Gaussian number plane: A number $z=a+\mathrm{i}\,b$ is represented by the point whose abscissa is a and ordinate is b **(Fig. 1.5)**. The real numbers are on the axis of abscissae which is also called the real axis, the pure imaginary numbers are on the axis of ordinates which is also called the imaginary axis. On this plane every point is given uniquely by its *position vector* or

radius vector (see 3.5.1.1, **6.**, p. 181), so every complex number corresponds to a vector which starts at the origin and is directed to the point defined by the complex number. So, complex numbers can be represented as points or as vectors **(Fig. 1.6)**.

1.5.2.2 Equality of Complex Numbers

Two complex numbers are equal by definition if their *real parts* and *imaginary parts* are equal to each other. From a geometric viewpoint, two complex numbers are equal if the position vectors corresponding to them are equal. In the opposite case the complex numbers are not equal. The notions "greater" and "smaller" are meaningless for complex numbers.

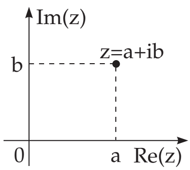


Figure 1.5

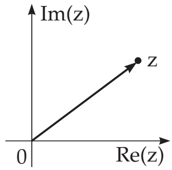


Figure 1.6

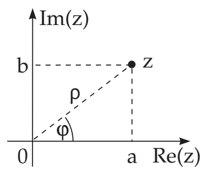


Figure 1.7

1.5.2.3 Trigonometric Form of Complex Numbers

The form

$$z = a + \mathrm{i}\, b \tag{1.130a}$$

is called the algebraic form of the complex number. Using polar coordinates yields the *trigonometric form of the complex numbers* **(Fig. 1.7)**:

$$z = \rho(\cos\varphi + \mathrm{i}\,\sin\varphi). \tag{1.130b}$$

The length of the position vector of a point $\rho = |z|$ is called the *absolute value* or the *magnitude of the complex number*, the angle φ, given in radian measure, is called the *argument of the complex number* and is denoted by $\arg z$:

$$\rho = |z|, \quad \varphi = \arg z = \omega + 2k\pi \ \text{ with } \ 0 \le \rho < \infty, \quad -\pi < \omega \le +\pi, \quad k = 0, \pm 1, \pm 2, \dots . \tag{1.130c}$$

One calls φ the *principal value of the argument of the complex number*.

The relations between ρ, φ and a, b for a point are the same as between the Cartesian and polar coordinates of a point (see 3.5.2.2, p. 192):

$$a = \rho\cos\varphi, \quad (1.131\text{a}) \qquad b = \rho\sin\varphi, \quad (1.131\text{b}) \qquad \rho = \sqrt{a^2 + b^2}, \quad (1.131\text{c})$$

$$\varphi = \begin{cases} \arccos\dfrac{a}{\rho} & \text{for } b \ge 0,\ \rho > 0, \\ -\arccos\dfrac{a}{\rho} & \text{for } b < 0,\ \rho > 0, \\ \text{undefined} & \text{for } \rho = 0 \end{cases} \quad (1.131\text{d}) \qquad \varphi = \begin{cases} \arctan\dfrac{b}{a} & \text{for } a > 0, \\ +\dfrac{\pi}{2} & \text{for } a = 0,\ b > 0, \\ -\dfrac{\pi}{2} & \text{for } a = 0,\ b < 0, \\ \arctan\dfrac{b}{a} + \pi & \text{for } a < 0,\ b \ge 0, \\ \arctan\dfrac{b}{a} - \pi & \text{for } a < 0,\ b < 0. \end{cases} \quad (1.131\text{e})$$

The complex number $z = 0$ has absolute value equal to zero; its argument $\arg 0$ is undefined.

1.5.2.4 Exponential Form of a Complex Number

The representation

$$z = \rho e^{\mathrm{i}\varphi} \tag{1.132a}$$

is called the *exponential form of the complex number*, where ρ is the magnitude and φ is the argument. The *Euler relation* is the formula

$$e^{\mathrm{i}\varphi} = \cos\varphi + \mathrm{i}\,\sin\varphi\,. \tag{1.132b}$$

■ Representation of a complex number in three forms:

a) $z = 1 + \mathrm{i}\sqrt{3}$ (algebraic form), **b)** $z = 2\left(\cos\frac{\pi}{3} + \mathrm{i}\,\sin\frac{\pi}{3}\right)$ (trigonometric form),

c) $z = 2\,e^{\mathrm{i}\frac{\pi}{3}}$ (exponential form), considering the principal value of it.
Without restriction to the principal value holds the representation

d) $z = 1+\mathrm{i}\sqrt{3} = 2\exp\left[\mathrm{i}\left(\frac{\pi}{3} + 2k\pi\right)\right] = 2\left[\cos\left(\frac{\pi}{3} + 2k\pi\right) + \mathrm{i}\,\sin\left(\frac{\pi}{3} + 2k\pi\right)\right] \quad (k = 0, \pm 1, \pm 2, \ldots)\,.$

1.5.2.5 Conjugate Complex Numbers

Two complex numbers z and z^* are called *conjugate complex numbers* if their real parts are equal and their imaginary parts differ only in sign:

$$\mathrm{Re}(z^*) = \mathrm{Re}(z)\,, \quad \mathrm{Im}(z^*) = -\mathrm{Im}(z)\,. \tag{1.133a}$$

The geometric interpretation of points corresponding to the conjugate complex numbers are points symmetric with respect to the real axis. Conjugate complex numbers have the same absolute value, their arguments differ only in sign:

$$z = a + \mathrm{i}\,b = \rho(\cos\varphi + \mathrm{i}\,\sin\varphi) = \rho e^{\mathrm{i}\varphi}, \tag{1.133b}$$

$$z^* = a - \mathrm{i}\,b = \rho(\cos\varphi - \mathrm{i}\,\sin\varphi) = \rho e^{-\mathrm{i}\varphi}. \tag{1.133c}$$

Instead of z^* one often uses the notation $\overline{z}$ for the conjugate of z.

1.5.3 Calculation with Complex Numbers

1.5.3.1 Addition and Subtraction

Addition and subtraction of two or more complex numbers given in algebraic form is defined by the formula

$$\begin{aligned} z_1 + z_2 - z_3 + \cdots &= (a_1 + \mathrm{i}\,b_1) + (a_2 + \mathrm{i}\,b_2) - (a_3 + \mathrm{i}\,b_3) + \cdots \\ &= (a_1 + a_2 - a_3 + \cdots) + \mathrm{i}\,(b_1 + b_2 - b_3 + \cdots)\,. \end{aligned} \tag{1.134}$$

The calculation can be done in the same way as doing with usual binomials. As a geometric interpretation of addition and subtraction can be considered the addition and subtraction of the corresponding vectors **(Fig. 1.8)**. For these the usual rules for vector calculations are to be used (see 3.5.1.1, p. 181). For z and z^*, $z + z^*$ is always real, and $z - z^*$ is pure imaginary.

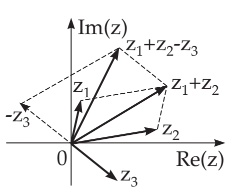


Figure 1.8

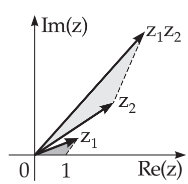


Figure 1.9

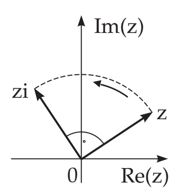


Figure 1.10

1.5.3.2 Multiplication

The multiplication of two complex numbers z_1 and z_2 given in algebraic form is defined by the following formula

$$z_1 z_2 = (a_1 + \mathrm{i}\, b_1)(a_2 + \mathrm{i}\, b_2) = (a_1 a_2 - b_1 b_2) + \mathrm{i}\,(a_1 b_2 + b_1 a_2)\,. \tag{1.135a}$$

For numbers given in trigonometric form holds

$$\begin{aligned} z_1 z_2 &= [\rho_1(\cos\varphi_1 + \mathrm{i}\,\sin\varphi_1)][\rho_2(\cos\varphi_2 + \mathrm{i}\,\sin\varphi_2)] \\ &= \rho_1\rho_2[\cos(\varphi_1+\varphi_2) + \mathrm{i}\,\sin(\varphi_1+\varphi_2)], \end{aligned} \tag{1.135b}$$

i.e., the absolute value of the product is equal to the product of the absolute values of the factors, and the argument of the product is equal to the sum of the arguments of the factors. The exponential form of the product is

$$z_1 z_2 = \rho_1\rho_2 e^{\mathrm{i}(\varphi_1+\varphi_2)}. \tag{1.135c}$$

The geometric interpretation of the product of two complex numbers z_1 and z_2 is a vector (**Fig. 1.9**). It is generated by rotation of the vector corresponding to z_1 by the argument of the vector z_2 (clockwise or counterclockwise according to the sign of this argument), and the length of the vector will be stretched by $|z_2|$.
The product $z_1 z_2$ can also be represented with similar triangles **(Fig. 1.9)**. The multiplication of a complex number z by i means a rotation by $\pi/2$ and the absolute value does not change **(Fig. 1.10)**.
For z and z^*:

$$zz^* = \rho^2 = |z|^2 = a^2 + b^2. \tag{1.136}$$

1.5.3.3 Division

Division is defined as the inverse operation of multiplication. For complex numbers given in algebraic form holds

$$\frac{z_1}{z_2} = \frac{a_1 + \mathrm{i}\, b_1}{a_2 + \mathrm{i}\, b_2} = \frac{a_1 a_2 + b_1 b_2}{{a_2}^2 + {b_2}^2} + \mathrm{i}\,\frac{a_2 b_1 - a_1 b_2}{{a_2}^2 + {b_2}^2}\,. \tag{1.137a}$$

For complex numbers given in trigonometric form holds

$$\frac{z_1}{z_2} = \frac{\rho_1(\cos\varphi_1 + \mathrm{i}\,\sin\varphi_1)}{\rho_2(\cos\varphi_2 + \mathrm{i}\,\sin\varphi_2)} = \frac{\rho_1}{\rho_2}[\cos(\varphi_1-\varphi_2) + \mathrm{i}\,\sin(\varphi_1-\varphi_2)]\,, \tag{1.137b}$$

i.e., the absolute value of the quotient is equal to the ratio of the absolute values of the dividend and the divisor; the argument of the quotient is equal to the difference of the arguments.
For the exponential form follows

$$\frac{z_1}{z_2} = \frac{\rho_1}{\rho_2} e^{\mathrm{i}(\varphi_1-\varphi_2)}. \tag{1.137c}$$

In the geometric representation the vector corresponding to z_1/z_2 can be generated by a rotation of the vector representing z_1 by $-\arg z_2$, and then by a contraction by $|z_2|$.
Remark: Division by zero is impossible.

1.5.3.4 General Rules for the Basic Operations

Calculations with complex numbers $z = a + \mathrm{i}\, b$ are to be done in the same way as doing with ordinary binomials, but considering $\mathrm{i}^2 = -1$. Dividing a complex number by a complex number first the imaginary part of the denominator has to be removed by multiplying the numerator and the denominator of the fraction by the complex conjugate of the divisor. This is possible because

$$(a + \mathrm{i}\, b)(a - \mathrm{i}\, b) = a^2 + b^2 \tag{1.138}$$

is a real number.

■ $$\frac{(3-4\mathrm{i})(-1+5\mathrm{i})^2}{1+3\mathrm{i}} + \frac{10+7\mathrm{i}}{5\mathrm{i}} = \frac{(3-4\mathrm{i})(1-10\mathrm{i}-25)}{1+3\mathrm{i}} + \frac{(10+7\mathrm{i})\mathrm{i}}{5\mathrm{i}\,\mathrm{i}} = \frac{-2(3-4\mathrm{i})(12+5\mathrm{i})}{1+3\mathrm{i}} +$$

$$\frac{7-10\mathrm{i}}{5} = \frac{-2(56-33\mathrm{i}\,)(1-3\mathrm{i}\,)}{(1+3\mathrm{i}\,)(1-3\mathrm{i}\,)}+\frac{7-10\mathrm{i}}{5} = \frac{-2(-43-201\mathrm{i}\,)}{10}+\frac{7-10\mathrm{i}}{5} = \frac{1}{5}(50+191\mathrm{i}\,) = 10+38.2\mathrm{i}\,.$$

1.5.3.5 Taking Powers of Complex Numbers

The n-th power of a complex number could be calculated using the binomial formula, but it would be very inconvenient. For practical reasons the trigonometric form is to be used and the so-called *de Moivre formula*:

$$[\rho(\cos\varphi + \mathrm{i}\,\sin\varphi)]^n = \rho^n(\cos n\varphi + \mathrm{i}\,\sin n\varphi)\,, \tag{1.139a}$$

i.e., the absolute value is raised to the n-th power, and the argument is multiplied by n. In particular, holds:

$$\mathrm{i}^{\,2} = -1, \quad \mathrm{i}^{\,3} = -\mathrm{i}\,, \quad \mathrm{i}^{\,4} = +1 \qquad (1.139\mathrm{b}) \qquad \text{in general} \qquad \mathrm{i}^{\,4n+k} = \mathrm{i}^{\,k}\,. \tag{1.139c}$$

1.5.3.6 Taking the n-th Root of a Complex Number

Taking of the n-th root is the inverse operation of taking powers. For $z = \rho(\cos\varphi + \mathrm{i}\,\sin\varphi) \neq 0$ the notation

$$z^{1/n} = \sqrt[n]{z} \quad (n > 0,\ \text{integer}), \tag{1.140a}$$

is the shorthand notation for the n different values

$$\omega_k = \sqrt[n]{\rho}\left(\cos\frac{\varphi + 2k\pi}{n} + \mathrm{i}\,\sin\frac{\varphi + 2k\pi}{n}\right),$$

$$(k = 0, 1, 2, \ldots, n-1). \tag{1.140b}$$

While addition, subtraction, multiplication, division, and taking a power with integer exponent have unique results, taking the n-th root has n different solutions ω_k.

The geometric interpretations of the points ω_k are the vertices of a regular n-gon whose center is at the origin. In **Fig. 1.11** the six values of $\sqrt[6]{z}$ are represented.

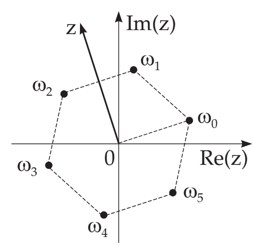


Figure 1.11

1.6 Algebraic and Transcendental Equations

1.6.1 Transforming Algebraic Equations to Normal Form

1.6.1.1 Definition

The variable x in the equality

$$F(x) = f(x) \tag{1.141}$$

is called the unknown if the equality is valid only for certain values $x_1, x_2, \ldots, x_n$ of the variable, and these values are called the *solutions* or the *roots* of the equation. Two equations are considered equivalent if they have exactly the same roots.

An equation is called an *algebraic equation* if the functions $F(x)$ and $f(x)$ are algebraic, i.e., they are rational or irrational expressions; of course one of them can be constant. Every algebraic equation can be transformed into the *normal form*

$$P(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 = 0 \tag{1.142}$$

by algebraic transformations. The roots of the original equation occur among the roots of the normal form, but under certain circumstances some are superfluous. The *leading coefficient* a_n is frequently transformed to the value 1.

The exponent n is called the *degree of the equation*.

■ Determine the normal form of the equation $\dfrac{x-1+\sqrt{x^2-6}}{3(x-2)} = 1 + \dfrac{x-3}{x}$. The transformations

step by step are:
$x(x-1+\sqrt{x^2-6}) = 3x(x-2)+3(x-2)(x-3)$, $x^2-x+x\sqrt{x^2-6} = 3x^2-6x+3x^2-15x+18$, $x\sqrt{x^2-6} = 5x^2-20x+18$, $x^2(x^2-6) = 25x^4-200x^3+580x^2-720x+324$, $24x^4-200x^3+586x^2-720x+324=0$. The result is an equation of fourth degree in normal form.

1.6.1.2 System of n Algebraic Equations

Every *system of algebraic equations* can be transformed to normal form, i.e., into a system of polynomial equations:

$$P_1(x,y,z,\ldots)=0\,,\ \ P_2(x,y,z,\ldots)=0\,,\ \ \ldots\,,\ \ P_n(x,y,z,\ldots)=0\,. \tag{1.143}$$

The P_i $(i=1,2,\ldots,n)$ are polynomials in $x, y, z, \ldots$.

■ Determine the normal form of the equation system: 1. $\frac{x}{\sqrt{y}}=\frac{1}{z}$, 2. $\frac{x-1}{y-1}=\sqrt{z}$, 3. $xy=z$.
The normal form is: 1. $x^2z^2-y=0$, 2. $x^2-2x+1-y^2z+2yz-z=0$, 3. $xy-z=0$.

1.6.1.3 Extraneous Roots

After transforming an algebraic equation into the normal form (1.142) it can happen that the equation $P(x)=0$ has some roots which are not solutions of the original equation (1.141). The roots of the equation $P(x)=0$ must be substituted into the original equation to check whether they are really solutions of (1.141).

Extraneous solutions can emerge if not invertible transformations are performed:

1. Vanishing denominator If the equation has the form

$$\frac{P(x)}{Q(x)}=0 \tag{1.144a}$$

with polynomials $P(x)$ and $Q(x)$, then the normal form of (1.144a) after multiplying by the denominator $Q(x)$ is:

$$P(x)=0. \tag{1.144b}$$

The roots of (1.144b) are the same as the roots of (1.144a), except the ones which are roots both of the numerator and of the denominator, i.e. which satisfy $P(x)=0$ and $Q(x)=0$. If $x=\alpha$ is a root of the denominator, then in the case $x=\alpha$ the multiplication by $Q(x)$ is a multiplication by zero. Every time when a non-identical transformation is performed, the checking of the solutions is necessary (see also 1.6.3.1, p. 43).

■ $\frac{x^3}{x-1}=\frac{1}{x-1}$. The corresponding normal form is $x^4-x^3-x+1=0$. $x_1=1$ is a solution of the normal form, but it is not a solution of the original equation, since the fractions are not defined for $x=1$.

2. Irrational equations If the original equation contains radicals, the normal form is usually achieved by powering. E.g. squaring is not an identical transformation (since it is not invertible).

■ $\sqrt{x+7}+1=2x$ or $\sqrt{x+7}=2x-1$. By squaring both sides of the second form of the equation its normal form is $4x^2-5x-6=0$, and the roots are $x_1=2$ and $x_2=-3/4$. The root $x_1=2$ is a solution of the original equation, but the root $x_2=-3/4$ is not.

1.6.2 Equations of Degree at Most Four

1.6.2.1 Equations of Degree One (Linear Equations)

1. Normal Form

$$ax+b=0\ \ (a\neq 0)\,. \tag{1.145}$$

2. Number of Solutions
There is a unique solution

$$x_1 = -\frac{b}{a}. \tag{1.146}$$

1.6.2.2 Equations of Degree Two (Quadratic Equations)

1. Normal Form

$$ax^2 + bx + c = 0 \quad (a \neq 0) \tag{1.147a}$$

or divided by a:

$$x^2 + px + q = 0. \tag{1.147b}$$

2. Number of Real Solutions of a Real Equation Depending on the sign of the discriminant

$$D = 4ac - b^2 \text{ for } (1.147a) \text{ or } D = q - \frac{p^2}{4} \text{ for } (1.147b), \tag{1.148}$$

holds:
- for $D < 0$, there are two real solutions (two real roots),
- for $D = 0$, there is one real solution (two coincident roots),
- for $D > 0$, there is no real solution (two complex roots).

3. Properties of the Roots of a Quadratic Equation If x_1 and x_2 are the roots of the quadratic equation (1.147a) or (1.147b), then the following equalities hold:

$$x_1 + x_2 = -\frac{b}{a} = -p, \quad x_1 \cdot x_2 = \frac{c}{a} = q. \tag{1.149}$$

4. Solution of Quadratic Equations
Method 1: Factorization of

$$ax^2 + bx + c = a(x - \alpha)(x - \beta) \quad (1.150a) \qquad \text{or} \quad x^2 + px + q = (x - \alpha)(x - \beta), \quad (1.150b)$$

if it is successful, immediately gives the roots

$$x_1 = \alpha, \quad x_2 = \beta. \tag{1.151}$$

■ $x^2 + x - 6 = 0$, $x^2 + x - 6 = (x + 3)(x - 2)$, $x_1 = -3$, $x_2 = 2$.

Method 2: Using the solution formula in the cases $D \leq 0$:
a) For (1.147a) the solutions are

$$x_{1,2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \quad (1.152a) \qquad \text{or} \quad x_{1,2} = \frac{-\frac{b}{2} \pm \sqrt{\left(\frac{b}{2}\right)^2 - ac}}{a}. \quad (1.152b)$$

If b is an even integer the second formula is to be used.
b) For (1.147b) the solutions are

$$x_{1,2} = -\frac{p}{2} \pm \sqrt{\frac{p^2}{4} - q}. \tag{1.153}$$

1.6.2.3 Equations of Degree Three (Cubic Equations)

1. Normal Form

$$ax^3 + bx^2 + cx + d = 0 \quad (a \neq 0) \tag{1.154a}$$

or after dividing by a and substituting $y = x + \frac{b}{3a}$ there is

$$y^3 + 3py + 2q = 0 \quad \text{or in reduced form} \quad y^3 + p^*y + q^* = 0, \tag{1.154b}$$

where

$$q^* = 2q = \frac{2b^3}{27a^3} - \frac{bc}{3a^2} + \frac{d}{a} \quad \text{and} \quad p^* = 3p = \frac{3ac - b^2}{3a^2}. \tag{1.154c}$$

2. Number of Real Solutions Depending on the sign of the discriminant

$$D = q^2 + p^3 \tag{1.155}$$

holds:
- for $D > 0$, one real solution (one real and two complex roots),
- for $D < 0$, three real solutions (three different real roots),
- for $D = 0$, one real solution (one real root with multiplicity three) in the case $p = q = 0$; or two real solutions (a single and a double real root) in the case $p^3 = -q^2 \neq 0$.

3. Properties of the Roots of a Cubic Equation If x_1, x_2, and x_3 are the roots of the cubic equation (1.154a), then the following equalities hold:

$$x_1 + x_2 + x_3 = -\frac{b}{a}, \quad x_1x_2 + x_1x_3 + x_2x_3 = \frac{c}{a}, \quad x_1x_2x_3 = -\frac{d}{a}. \tag{1.156}$$

4. Solution of a Cubic Equation

Method 1: If it is possible to decompose the left-hand side into a product of linear terms

$$ax^3 + bx^2 + cx + d = a(x - \alpha)(x - \beta)(x - \gamma) \tag{1.157a}$$

one immediately gets the roots

$$x_1 = \alpha, \quad x_2 = \beta, \quad x_3 = \gamma. \tag{1.157b}$$

■ $x^3 + x^2 - 6x = 0$, $x^3 + x^2 - 6x = x(x + 3)(x - 2)$; $x_1 = 0$, $x_2 = -3$, $x_3 = 2$.

Method 2: Using the Formula of Cardano. By substituting $y = u + v$ the equation (1.154b) has the form

$$u^3 + v^3 + (u + v)(3uv + 3p) + 2q = 0. \tag{1.158a}$$

This equation is obviously satisfied if

$$u^3 + v^3 = -2q \quad \text{and} \quad uv = -p \tag{1.158b}$$

hold. Writing (1.158b) in the form

$$u^3 + v^3 = -2q, \qquad u^3v^3 = -p^3, \tag{1.158c}$$

there are two unknowns u^3 and v^3, the sum and product of which are known. Therefore using the Vieta root theorem (see 1.6.3.1, **3.**, p. 44) the solutions of the quadratic equation

$$w^2 - (u^3 + v^3)w + u^3v^3 = w^2 + 2qw - p^3 = 0 \tag{1.158d}$$

can be calculated:

$$w_1 = u^3 = -q + \sqrt{q^2 + p^3}, \qquad w_2 = v^3 = -q - \sqrt{q^2 + p^3}, \tag{1.158e}$$

so for the solution y of (1.154b) the *Cardano formula* results in

$$y = u + v = \sqrt[3]{-q + \sqrt{q^2 + p^3}} + \sqrt[3]{-q - \sqrt{q^2 + p^3}}. \tag{1.158f}$$

Since the third root of a complex number means three different numbers (see (1.140b), p. 38) there are nine different cases, but because of $uv = -p$, the solutions are reduced to the following three:

$$y_1 = u_1 + v_1 \text{ (if possible, consider the real third roots } u_1 \text{ and } v_1 \text{ such that } u_1v_1 = -p\text{)}, \tag{1.158g}$$

$$y_2 = u_1\left(-\frac{1}{2} + \frac{\mathrm{i}}{2}\sqrt{3}\right) + v_1\left(-\frac{1}{2} - \frac{\mathrm{i}}{2}\sqrt{3}\right), \tag{1.158h}$$

$$y_3 = u_1\left(-\frac{1}{2} - \frac{\mathrm{i}}{2}\sqrt{3}\right) + v_1\left(-\frac{1}{2} + \frac{\mathrm{i}}{2}\sqrt{3}\right). \tag{1.158i}$$

■ $y^3 + 6y + 2 = 0$ with $p = 2,\ q = 1$ and $q^2 + p^3 = 9$ and $u = \sqrt[3]{-1+3} = \sqrt[3]{2} = 1.2599$, $v = \sqrt[3]{-1-3} = \sqrt[3]{-4} = -1.5874$. The real root is $y_1 = u + v = -0.3275$, the complex roots are $y_{2,3} = -\frac{1}{2}(u+v) \pm \mathrm{i}\frac{\sqrt{3}}{2}(u-v) = 0.1638 \pm \mathrm{i} \cdot 2.4659$.

Method 3: For a real equation, the *auxiliary values* given in **Table 1.3** can be used. With p from (1.154c)

$$r = \pm\sqrt{|p|} \tag{1.159}$$

is substituted where the sign of r is the same as the sign of q. Next, using **Table 1.3**, one can determine the value of the auxiliary variable φ and with it the roots y_1, y_2 and y_3 depending on the signs of p and $D = q^2 + p^3$.

Table 1.3 Auxiliary values for the solution of equations of degree three

$p < 0$		$p > 0$
$q^2 + p^3 \le 0$	$q^2 + p^3 > 0$	
$\cos\varphi = \frac{q}{r^3}$	$\cosh\varphi = \frac{q}{r^3}$	$\sinh\varphi = \frac{q}{r^3}$
$y_1 = -2r\cos\frac{\varphi}{3}$ $y_2 = +2r\cos\left(60^\circ - \frac{\varphi}{3}\right)$ $y_3 = +2r\cos\left(60^\circ + \frac{\varphi}{3}\right)$	$y_1 = -2r\cosh\frac{\varphi}{3}$ $y_2 = r\cosh\frac{\varphi}{3} + \mathrm{i}\sqrt{3}\,r\sinh\frac{\varphi}{3}$ $y_3 = r\cosh\frac{\varphi}{3} - \mathrm{i}\sqrt{3}\,r\sinh\frac{\varphi}{3}$	$y_1 = -2r\sinh\frac{\varphi}{3}$ $y_2 = r\sinh\frac{\varphi}{3} + \mathrm{i}\sqrt{3}\,r\cosh\frac{\varphi}{3}$ $y_3 = r\sinh\frac{\varphi}{3} - \mathrm{i}\sqrt{3}\,r\cosh\frac{\varphi}{3}$

■ $y^3 - 9y + 4 = 0$. $p = -3,\ q = 2,\ q^2 + p^3 < 0,\ r = \sqrt{3},\ \cos\varphi = \frac{2}{3\sqrt{3}} = 0.3849,\ \varphi = 67°22'$.
$y_1 = -2\sqrt{3}\cos 22°27' = -3.201,\ y_2 = 2\sqrt{3}\cos(60° - 22°27') = 2.747,\ y_3 = 2\sqrt{3}\cos(60° + 22°27') = 0.455$.
Checking: $y_1 + y_2 + y_3 = 0.001$ which can be considered 0 for the accuracy of our calculations.

Method 4: Numerical approximate solution, see 19.1.2, p. 952; numerical approximate solution by the help of a nomogram, see 2.19, p. 128.

1.6.2.4 Equations of Degree Four

1. Normal Form

$$ax^4 + bx^3 + cx^2 + dx + e = 0 \quad (a \ne 0)\,. \tag{1.160}$$

If all the coefficients are real, this equation has 0 or 2 or 4 real solutions.

2. Special Forms If $b = d = 0$ holds, the roots of the *biquadratic equation*

$$ax^4 + cx^2 + e = 0 \tag{1.161a}$$

can be calculated by the formulas

$$x_{1,2,3,4} = \pm\sqrt{y}, \quad y = \frac{-c \pm \sqrt{c^2 - 4ae}}{2a}. \tag{1.161b}$$

For $a = e$ and $b = d$, the roots of the equation

$$ax^4 + bx^3 + cx^2 + bx + a = 0 \tag{1.161c}$$

can be calculated by the formulas

$$x_{1,2,3,4} = \frac{y \pm \sqrt{y^2 - 4}}{2}, \quad y = \frac{-b \pm \sqrt{b^2 - 4ac + 8a^2}}{2a}. \tag{1.161d}$$

3. Solution of a General Equation of Degree Four

Method 1: If somehow the left-hand side of the equation can be factorized

$$ax^4 + bx^3 + cx^2 + dx + e = 0 = a(x-\alpha)(x-\beta)(x-\gamma)(x-\delta) \tag{1.162a}$$

then the roots can be immediately determined:

$$x_1 = \alpha\,, \quad x_2 = \beta\,, \quad x_3 = \gamma\,, \quad x_4 = \delta. \tag{1.162b}$$

■ $x^4 - 2x^3 - x^2 + 2x = 0,\ x(x^2-1)(x-2) = x(x-1)(x+1)(x-2);$
$x_1 = 0,\ x_2 = 1,\ x_3 = -1,\ x_4 = 2.$

Method 2: The roots of the equation (1.162a) for $a = 1$ coincide with the roots of the equation

$$x^2 + (b+A)\frac{x}{2} + \left(y + \frac{by-d}{A}\right) = 0, \tag{1.163a}$$

where $A = \pm\sqrt{8y + b^2 - 4c}$ and y is one of the real roots of the equation of third degree

$$8y^3 - 4cy^2 + (2bd - 8e)y + e(4c - b^2) - d^2 = 0 \tag{1.163b}$$

with $B = \dfrac{b^3}{8} - \dfrac{bc}{2} \neq 0$. The case $B = 0$ gives by the help of the substitution $x = u - \dfrac{b}{4}$ a biquadratic equation of the form (1.161a) for u with $a = 1$.

Method 3: Approximate solution, see 19.1.2, p. 952.

1.6.2.5 Equations of Higher Degree

It is impossible to give a formula or a finite sequence of formulas which produce the roots of an equation of degree five or higher (see also 19.1.2.2,**2.**, p. 954).

1.6.3 Equations of Degree n

1.6.3.1 General Properties of Algebraic Equations

1. Roots

The left-hand side of the equation

$$x^n + a_{n-1}x^{n-1} + \ldots + a_0 = 0 \tag{1.164a}$$

is a polynomial $P_n(x)$ of degree n, and a solution of (1.164a) is a root of the polynomial $P_n(x)$. If α is a root of the polynomial, then $P_n(x)$ is divisible by $(x-\alpha)$. Generally

$$P_n(x) = (x-\alpha)P_{n-1}(x) + P_n(\alpha). \tag{1.164b}$$

Here $P_{n-1}(x)$ is a polynomial of degree $n-1$. If $P_n(x)$ is divisible by $(x-\alpha)^k$, but it is not divisible by $(x-\alpha)^{k+1}$ then α is called a *root of order* k of the equation $P_n(x) = 0$. In this case α is a common root of the polynomial $P_n(x)$ and its derivatives to order $(k-1)$.

2. Fundamental Theorem of Algebra

Every equation of degree n whose coefficients are real or complex numbers has n real or complex roots, where the roots of higher order are counted by their multiplicity. Denoting the roots of $P(x)$ by $\alpha, \beta, \gamma, \ldots$ and they have multiplicity $k, l, m, \ldots$, then the *product representation of the polynomial* is

$$P(x) = (x-\alpha)^k(x-\beta)^l(x-\gamma)^m \ldots\,. \tag{1.165a}$$

The solution of the equation $P(x) = 0$ can be simplified by reducing the equation to another one, which has the same roots, but only with multiplicity one (if possible). In order to get this, the polynomial is to be composed into a product of two factors

$$P(x) = Q(x)T(x), \tag{1.165b}$$

such that

$$T(x) = (x-\alpha)^{k-1}(x-\beta)^{l-1} \ldots, \quad Q(x) = (x-\alpha)(x-\beta) \ldots\,. \tag{1.165c}$$

Because the roots of the polynomial $P(x)$ with higher multiplicity are the roots of its derivative $P'(x)$, too, $T(x)$ is the greatest common devisor of the polynomial $P(x)$ and its derivative $P'(x)$ (see 1.1.6.5, p.14). Dividing $P(x)$ by $T(x)$ yields the polynomial $Q(x)$ which has all the roots of $P(x)$, and each root occurs with multiplicity one.

3. Theorem of Vieta About Roots

The relations between the n roots $x_1, x_2, \ldots, x_n$ and the coefficients of the equation (1.164a) are:

$$\begin{aligned} x_1 + x_2 + \ldots + x_n &= \sum_{i=1}^{n} x_i = -a_{n-1}, \\ x_1x_2 + x_1x_3 + \ldots + x_{n-1}x_n &= \sum_{\substack{i,j=1\\ i<j}}^{n} x_ix_j = a_{n-2}, \\ x_1x_2x_3 + x_1x_2x_4 + \ldots + x_{n-2}x_{n-1}x_n &= \sum_{\substack{i,j,k=1\\ i<j<k}}^{n} x_ix_jx_k = -a_{n-3}, \\ &\ldots \\ x_1x_2\ldots x_n &= (-1)^n a_0. \end{aligned} \tag{1.166}$$

1.6.3.2 Equations with Real Coefficients

1. Complex Roots

Polynomial equations with real coefficients can also have complex roots but only pairwise conjugate complex numbers, i.e., if $\alpha = a + \mathrm{i}\,b$ is a root, then $\beta = a - \mathrm{i}\,b$ is also a root, and it has the same multiplicity. The expressions $p = -(\alpha + \beta) = -2a$ and $q = \alpha\beta = a^2 + b^2$ satisfy the unequation $\left(\frac{p}{2}\right)^2 - q < 0$, so that

$$(x - \alpha)(x - \beta) = x^2 + px + q \tag{1.167}$$

holds. Substituting the product corresponding to (1.167) for every pair of factors in (1.165a), one gets a decomposition of the polynomial with real coefficients into *real factors*.

$$\begin{aligned} P(x) = {} & (x - \alpha_1)^{k_1}(x - \alpha_2)^{k_2} \cdots (x - \alpha_l)^{k_l} \\ & \cdot(x^2 + p_1x + q_1)^{m_1}(x^2 + p_2x + q_2)^{m_2} \cdots (x^2 + p_rx + q_r)^{m_r}. \end{aligned} \tag{1.168}$$

Here $\alpha_1, \alpha_2, \ldots, \alpha_l$ are the l real roots of the polynomial $P(x)$. It also has r pairs of conjugate complex roots, which are the roots of the quadratic factors $x^2 + p_ix + q_i$ $(i = 1, 2, \ldots, r)$. The numbers α_j $(j = 1, 2, \ldots, l)$, p_i and q_i $(i = 1, 2, \ldots, r)$ are real and the inequalities $\left(\frac{p_i}{2}\right)^2 - q_i < 0$ hold.

2. Number of Roots of an Equation with Real Coefficients

According to (1.167) every equation of odd degree has at least one real root. The number of further real roots of (1.164a) between two arbitrary real numbers $a < b$, can be determined in the following way:

a) Separate the Multiple Roots: Separating the multiple roots of $P(x) = 0$, yields an equation which has all the roots of the original equation, but only with multiplicity one. Then the form mentioned in the case of the fundamental theorem must be produced.
For practical reasons it is a good idea to start with the determination of the *Sturm chain* (the *Sturm functions* (1.169)). This is almost the same as the Euclidean algorithm for determining the greatest common devisor, but it gives some further information. If P_m is not a constant then $P(x)$ has multiple roots, which must be separated. Therefore in the following it can be assumed that $P(x) = 0$ has no multiple roots.

b) Creating the Sequence of Sturm Functions:

$$P(x), P'(x), P_1(x), P_2(x), \ldots, P_m = \text{const.} \tag{1.169}$$

Here $P(x)$ is the left-hand side of the equation, $P'(x)$ is the first derivative of $P(x)$, $P_1(x)$ is the remainder on division of $P(x)$ by $P'(x)$, but with the opposite sign, $P_2(x)$ is the remainder on division of $P'(x)$ by $P_1(x)$ similarly with the opposite sign, etc.; $P_m = \text{const}$ is the last non-zero remainder, but it must be a constant, otherwise $P(x)$ and $P'(x)$ have common devisors, and $P(x)$ has multiple roots. In order to simplify the calculations the remainders can be multiplied by positive numbers, what does not change the result.

c) Theorem of Sturm: If A is the number of changes in sign, i.e. the number of changes from "+" to "−" and vice versa, in the sequence (1.169) for $x = a$, and B is the number of changes in sign in the sequence (1.169) for $x = b$, then the difference $A - B$ is equal to the number of real roots of $P(x) = 0$ in the interval $[a, b]$. If in the sequence some numbers are equal to zero, then they should not be considered in the sign change count.

■ Determination of the number of roots of the equation $x^4 - 5x^2 + 8x - 8 = 0$ in the interval $[0, 2]$. The calculations by the *Sturm functions* are: $P(x) = x^4 - 5x^2 + 8x - 8$; $P'(x) = 4x^3 - 10x + 8$; $P_1(x) = 5x^2 - 12x + 16$; $P_2(x) = -3x + 284$; $P_3 = -1$. Substituting $x = 0$ results in the sequence $-8, +8, +16, +284, -1$ with two changes in sign, substituting $x = 2$ results in $+4, +20, +12, +278, -1$ with one change in sign, so $A - B = 2 - 1 = 1$, i.e., between 0 and 2 there is one root.

d) Descartes Rule: The number of positive roots of the equation $P(x) = 0$ is not greater than the number of changes of sign in the sequence of coefficients of the polynomial $P(x)$, and these two numbers can differ from each other only by an even number.

■ What can be told about the roots of the equation $x^4 + 2x^3 - x^2 + 5x - 1 = 0$? The coefficients in the equation have signs $+\,,+\,,-\,,+\,,-\,$, i.e., there are three changes of sign. By the rule of Descartes the equation has either three or one roots. Because on replacing x by $-x$ the roots of the equation change their signs, and on replacing x by $x + h$ the roots are shifted by h, the number of negative roots, or the roots greater than h can be estimated by the help of the rule of Descartes. In the given example replacing x by $-x$ yields $x^4 - 2x^3 - x^2 - 5x - 1 = 0$, i.e., the equation has at most one negative root. Replacing x by $x + 1$ yields $x^4 + 6x^3 + 11x^2 + 13x + 6 = 0$, i.e., every positive root of the equation (one or three) is smaller than 1.

3. Solution of Equations of Degree n

Usually equations with $n > 4$ can be solved only approximately. In practice, approximate methods are also used to get solutions of equations of degree three or four (see 19.1.2.3, p. 954).
In order to determine certain real roots of an algebraic equation the general numerical procedures for non-linear equations can be used (see 19.1, p. 949). In order to determine all roots, including the complex roots of an algebraic equation of degree n the Brodetsky-Smeal method can be used (see [1.7], [19.31]). In order to determine complex roots one can use the Bairstow method (see [19.31]).

1.6.4 Reducing Transcendental Equations to Algebraic Equations

1.6.4.1 Definition

An equation $F(x) = f(x)$ is transcendental if at least one of the functions $F(x)$ or $f(x)$ is not algebraic.
■ **A:** $3^x = 4^{x-2} \cdot 2^x$, ■ **B:** $2\log_5(3x - 1) - \log_5(12x + 1) = 0$, ■ **C:** $3\cosh x = \sinh x + 9$,

■ **D:** $2^{x-1} = 8^{x-2} - 4^{x-2}$, ■ **E:** $\sin x = \cos^2 x - \frac{1}{4}$, ■ **F:** $x \cos x = \sin x$.

In some cases it is possible to reduce the solution of a transcendental equation to the solution of an algebraic equation, for instance by appropriate substitutions. In general, transcendental equations can be solved only approximately. In the following sections some special transcendental equations are

discussed which can be reduced to algebraic equations.

1.6.4.2 Exponential Equations

Exponential equations can be reduced to algebraic equations in the following two cases, if the unknown x or a polynomial $P(x)$ is only in the exponent of some quantities $a, b, c, \ldots$:

a) If the powers $a^{P_1(x)}$, $b^{P_2(x)}, \ldots$ are connected by multiplication or division, then the logarithm can be taken on an arbitrary base.

■ $3^x = 4^{x-2} \cdot 2^x$; $x \log 3 = (x-2)\log 4 + x \log 2$; $x = \dfrac{2\log 4}{\log 4 - \log 3 + \log 2}$.

b) If $a, b, c, \ldots$ are integer (or rational) powers of the same number k, i.e., $a = k^n$, $b = k^m$, $c = k^l, \ldots$, holds, then by substituting $y = k^x$ one can get an algebraic equation for y, and after solving it follows the solution $x = \dfrac{\log y}{\log k}$.

■ $2^{x-1} = 8^{x-2} - 4^{x-2}$; $\dfrac{2^x}{2} = \dfrac{2^{3x}}{64} - \dfrac{2^{2x}}{16}$. Substitution of $y = 2^x$ results in $y^3 - 4y^2 - 32y = 0$ and $y_1 = 8$, $y_2 = -4$, $y_3 = 0$; $2^{x_1} = 8$, $2^{x_2} = -4$, $2^{x_3} = 0$, so $x_1 = 3$ follows. There are no further real roots.

1.6.4.3 Logarithmic Equations

Logarithmic equations can be reduced to algebraic equations in the following two cases, if the unknown x or a polynomial $P(x)$ is only under the logarithm sign:

a) If the equation contains only the logarithm of the same expression, then by introducing this as a new unknown, one can solve the equation with respect to it. The original unknown can be determined by using the logarithm.

■ $m[\log_a P(x)]^2 + n = a\sqrt{[\log_a P(x)]^2 + b}$. The substitution $y = \log_a P(x)$ yields the equation $my^2 + n = a\sqrt{y^2 + b}$. After solving for y one gets the solution for x from the equation $P(x) = a^y$.

b) If the equation is a linear combination of logarithms of polynomials of x, on the same base a, with integer coefficients $m, n, \ldots$, i.e., it has the form $m \log_a P_1(x) + n \log_a P_2(x) + \ldots = 0$, then the left-hand side can be written as the logarithm of a rational expression. (The original equation may contain rational coefficients and rational expressions under the logarithm, or logarithms with different bases, if the bases are rational powers of each other.)

■ $2\log_5(3x-1) - \log_5(12x+1) = 0$, $\log_5 \dfrac{(3x-1)^2}{12x+1} = \log_5 1$, $\dfrac{(3x-1)^2}{12x+1} = 1$; $x_1 = 0$, $x_2 = 2$. Substituting $x_1 = 0$ in the original equation gives negative values in the logarithm, i.e., this logarithm is a complex value, so $x = 0$ is not a solution.

1.6.4.4 Trigonometric Equations

Trigonometric equations can be reduced to algebraic equations if the unknown x or the expression $nx+a$ with integer n is only in the argument of the trigonometric functions. After using the trigonometric formulas (see 2.7.2, p.81) the equation will contain only one unique function containing x, and after replacing it by y an algebraic equation arises. The solution for x is obtained from the solutions for y, naturally taking the multi-valuedness of the solution into consideration.

■ $\sin x = \cos^2 x - \dfrac{1}{4}$ or $\sin x = 1 - \sin^2 x - \dfrac{1}{4}$. Substituting $y = \sin x$ yields $y^2 + y - \dfrac{3}{4} = 0$ and $y_1 = \dfrac{1}{2}$, $y_2 = -\dfrac{3}{2}$. The result y_2 gives no real solution, because $|\sin x| \le 1$ for all real x; from $y_1 = \dfrac{1}{2}$

follows $x = \dfrac{\pi}{6} + 2k\pi$ and $x = \dfrac{5\pi}{6} + 2k\pi$ with $k = 1, 2, 3, \ldots$.

1.6.4.5 Equations with Hyperbolic Functions

Equations with hyperbolic functions can be reduced to algebraic equations if the unknown x is only in the argument of the hyperbolic functions. Rewriting the hyperbolic functions as exponential expressions, then substituting $y = e^x$ and $\dfrac{1}{y} = e^{-x}$, and the result is an algebraic equation for y. After solving this the solution is $x = \ln y$.

■ $3\cosh x = \sinh x + 9$; $\dfrac{3(e^x + e^{-x})}{2} = \dfrac{e^x - e^{-x}}{2} + 9$; $e^x + 2e^{-x} - 9 = 0$; $y + \dfrac{2}{y} - 9 = 0$, $y^2 - 9y + 2 = 0$; $y_{1,2} = \dfrac{9 \pm \sqrt{73}}{2}$; $x_1 = \ln\dfrac{9 + \sqrt{73}}{2} \approx 2.1716$, $x_2 = \ln\dfrac{9 - \sqrt{73}}{2} \approx -1.4784$.

2 Functions

2.1 Notion of Functions

2.1.1 Definition of a Function

2.1.1.1 Function

If x and y are two variable quantities, and if there is a rule which assigns a unique value of y to a given value of x, then y is called a function of x, using the notation

$$y = f(x)\,. \tag{2.1}$$

The variable x is called the *independent variable* or the *argument* of the function y. The values of x, to which a value of y is assigned, form the *domain* D of the function $f(x)$. The variable y is called the *dependent variable*; the values of y form the *range* W of the function $f(x)$. Functions can be represented by the points (x, y) as curves, or graphs of the function.

2.1.1.2 Real Functions

If both the domain and the range contain only real numbers the function $y = f(x)$ is called a *real function* of a *real variable*.

■ **A:** $y = x^2$ with $D:\ -\infty < x < \infty,\ W:\ 0 \le y < \infty$.

■ **B:** $y = \sqrt{x}$ with $D:\ 0 \le x < \infty,\quad W:\ 0 \le y < \infty$.

2.1.1.3 Functions of Several Variables

If the variable y depends on several independent variables $x_1, x_2, \ldots, x_n$, then the notation

$$y = f(x_1, x_2, \ldots, x_n) \tag{2.2}$$

is used for a function of several variables (see 2.18, p. 118).

2.1.1.4 Complex Functions

If the dependent and independent variables are *complex numbers w and z* respectively, then $w = f(z)$ means a *complex function* of a *complex variable*, (see 14.1, p. 731). *Complex-valued functions* $w(x)$ are called complex functions even if they have real arguments x.

2.1.1.5 Further Functions

In different fields of mathematics, for instance in vector analysis and in vector field theory (see 13.1, p. 701), other types of functions are to be considered whose arguments and values are defined as follows:

1. The arguments are real – the function values are vectors.

■ **A:** Vector functions (see 13.1.1, p. 701).

■ **B:** Parameter representations of curves (see 3.6.2, p. 256).

2. The arguments are vectors – the function values are real numbers.

■ Scalar fields (see 13.1.2, p. 702).

3. The arguments are vectors – the function values are vectors.

■ **A:** Vector fields (see 13.1.3, p. 704). ■ **B:** Parametric representations or vector forms of surfaces (see 3.6.3, p. 261).

2.1.1.6 Functionals

If a real number is assigned to every function $x = x(t)$ of a given class of functions, then it is called a *functional*.

■ **A:** If $x(t)$ is a given function which is integrable on $[a, b]$, then $f(x) = \int_a^b x(t)\,dt$ is a linear functional defined on the set of continuous functions x integrable on $[a, b]$ (see 12.5, p. 677).

I.N. Bronshtein et al., *Handbook of Mathematics*,
DOI 10.1007/978-3-662-46221-8_2

■ **B:** Integral expressions in variational problems (see 10.1, p. 610).

2.1.1.7 Functions and Mappings

Suppose there are given two non-empty sets X and Y. A *mapping*, which is denoted by

$$f: X \to Y, \tag{2.3}$$

is a rule, by which a uniquely defined element y of Y is assigned to every element x of X. The element y is called the *image* of x, as formula $y = f(x)$. The set Y is called the *image space* or *range* of f, the set X is called the *original space* or *domain* of f.

■ **A:** If both the original and the image spaces are subsets of real numbers, i.e., $X = D \subset \mathbb{R}$ and $Y = W \subset \mathbb{R}$ hold, then (2.3) defines a real function $y = f(x)$ of the real variable x.

■ **B:** If f is a matrix $\mathbf{A} = (a_{ij})$ $(i = 1, 2, \ldots, m; j = 1, 2, \ldots, n)$ of type (m, n) and $X = \mathbb{R}^n$ and $Y = \mathbb{R}^m$, then (2.3) defines a mapping from $\mathbb{R}^n$ into $\mathbb{R}^m$. The rule (2.3) is given by the following system of m linear equations:

$$\underline{\mathbf{y}} = \mathbf{A}\underline{\mathbf{x}} \qquad \text{or} \qquad \begin{array}{l} y_1 = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \\ y_2 = a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ y_m = a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n \end{array},$$

i.e. $\mathbf{A}\underline{\mathbf{x}}$ means the product of the matrix $\mathbf{A}$ and the vector $\underline{\mathbf{x}}$.

Remarks:

1. The notion of mapping is a generalization of the notion of function. So, some mappings are sometimes called functions.

2. The important properties of mappings can be found in 5.2.3, **5.** p. 333.

3. A mapping, which assigns to every element from an abstract space X a unique element usually from a different abstract space Y, is called an *operator*. Here an *abstract space* usually means a function space, since the most important spaces in applications consist of functions. Abstract spaces are for instance linear spaces (see vector spaces 5.3.8, p. 365), metric spaces (see 12.2, p. 662) and normed spaces (see 12.3, p. 669).

2.1.2 Methods for Defining a Real Function

2.1.2.1 Defining a Function

A function can be defined in several different ways, for instance by a table of values, by graphical representation, i.e., by a curve, by a formula, which is called an *analytic expression*, or piece by piece with different formulas. Only such values of the independent variable can belong to the *domain* of an analytic expression for which the function makes sense, i.e., it takes a unique, finite real value. If the domain is not otherwise defined, the domain is considered as the maximal set for which the definition makes sense.

2.1.2.2 Analytic Representation of a Function

Usually the following three forms are in use:

1. Explicit Form:

$$y = f(x). \tag{2.4}$$

■ $y = \sqrt{1 - x^2}, \quad -1 \le x \le 1, \quad y \ge 0$. Here the graph is the upper half of the unit circle centered at the origin.

2. Implicit Form:

$$F(x, y) = 0, \tag{2.5}$$

in the case when there is a unique y which satisfies this equation, or it can be told which solution is considered to be the value of the function.

■ $x^2 + y^2 - 1 = 0, \quad -1 \le x \le +1, \quad y \ge 0$. Here the graph is again the upper half of the unit

circle centered at the origin. It should be emphasized that $x^2 + y^2 + 1 = 0$ itself does not define a real function.

3. Parametric Form:

$$x = \varphi(t), \quad y = \psi(t). \tag{2.6}$$

The corresponding values of x and y are given as functions of an auxiliary variable t, which is called a *parameter*. The functions $\varphi(t)$ and $\psi(t)$ must have the same domain. This representation defines a real function only if $x = \varphi(t)$ defines a one-to-one correspondence between x and t.

■ $x = \varphi(t), \quad y = \psi(t)$ with $\varphi(t) = \cos t$ and $\psi(t) = \sin t, \quad 0 \le t \le \pi$. Here the graph is again the upper half of the unit circle centered at the origin.

Remark: Functions given in parametric form sometimes do not have any explicit or implicit parameter-free equation.

■ $x = t + 2\sin t = \varphi(t)$, $y = t - \cos t = \psi(t)$.

Examples for Functions Given Piece by Piece:

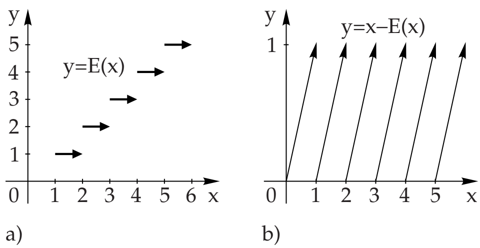


Figure 2.1

■ **A:** $y = E(x) = \text{int}(x) = [x] = n$ for $n \le x < n+1, \ n$ integer.
The function $E(x)$ or $\text{int}(x)$ (read "*integer part of* x") means the greatest integer less than or equal to x.

■ **B:** The function $y = \text{frac}(x) = x - [x]$ (read "*fractional part of* x") gives the difference of x and $[x]$ **(Fig. 2.1b)**. **Fig. 2.1a,b** shows the corresponding graphical representations, where the arrow-heads mean that the endpoints do not belong to the curves.

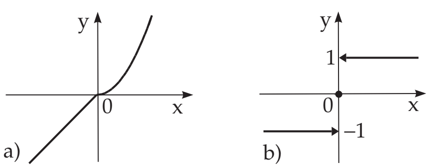


Figure 2.2

■ **C:** $y = \begin{cases} x & \text{for } x \le 0, \\ x^2 & \text{for } x \ge 0, \end{cases}$ **(Fig. 2.2a)**.

■ **D:** $y = \text{sign}(x) = \begin{cases} -1 & \text{for } x < 0, \\ 0 & \text{for } x = 0, \\ +1 & \text{for } x > 0, \end{cases}$

(Fig. 2.2b). By $\text{sign}(x)$ (read "signum x"), the *sign function* is denoted.

2.1.3 Certain Types of Functions

2.1.3.1 Monotone Functions

If a function satisfies the relations

$$f(x_2) \ge f(x_1) \quad \text{or} \quad f(x_2) \le f(x_1), \tag{2.7a}$$

for arbitrary arguments x_1 and x_2 with $x_2 > x_1$ in its domain, then it is called *monotonically increasing* or *monotonically decreasing* **(Fig. 2.3a,b)**.

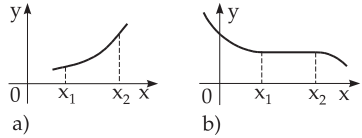


Figure 2.3

If one of the above relations (2.7a) does not hold for every x in the domain of the function, but it is valid, e.g., in an interval or on a half-axis, then the function is called *monotonic in this domain.* Functions satisfying the relations

$$f(x_2) > f(x_1) \quad \text{or} \quad f(x_2) < f(x_1), \tag{2.2}$$

i.e., when the equality never holds in (2.7a), are called *strictly monotonically increasing* or *strictly monotonically decreasing*. In **Fig. 2.3a** there is a representation of a strictly monotonically increasing function; in **Fig. 2.3b** there is the graph of a monotonically decreasing function being constant between x_1 and x_2.

■ $y = e^{-x}$ is strictly monotonically decreasing, $y = \ln x$ is strictly monotonically increasing.

2.1.3.2 Bounded Functions

A function is called *bounded above* if there is a number (called an *upper bound*) such that the values of the function never exceed it. A function is called *bounded below* if there is a number (called a *lower bound*) such that the values of the function are never less than this number. If a function is bounded above and below, it simply is called *bounded.*

■ **A:** $y = 1 - x^2$ is bounded above ($y \le 1$). ■ **B:** $y = e^x$ is bounded below ($y > 0$).

■ **C:** $y = \sin x$ is bounded ($-1 \le y \le +1$). ■ **D:** $y = \dfrac{4}{1+x^2}$ is bounded ($0 < y \le 4$).

2.1.3.3 Extreme Values of Functions

The function $f(x)$ with domain D has an *absolute* or *global maximum* at the point a, if for all $x \in D$

$$f(a) \ge f(x) \tag{2.8a}$$

holds. The function $f(x)$ has a *relative* or *local maximum* at the point a, if the inequality (2.8a) holds only in an environment of the point a, i.e. for all x with $a - \varepsilon < x < a + \varepsilon, \varepsilon > 0, x \in D$.

In analogy the definition of an *absolute* or *global minimum* as well as for a *relative* or *local minimum* can be given, but the inequality (2.8a) is to be replaced by

$$f(a) \le f(x)\,. \tag{2.8b}$$

Remarks:

a) The notions maximum and minimum, are called the *extreme values*, they are not coupled to the differentiability of functions, i.e., they hold also for functions which are not differentiable in some points of the domain. Examples are discontinuities of curves (see **Figs.2.9**, p. 58 and **6.10b,c**, p. 443).

b) Criterions for the determination of extreme values of differentiable functions see in 6.1.5.2, p. 443.

2.1.3.4 Even Functions

Even functions **(Fig. 2.4a)** satisfy the relation

$$f(-x) = f(x). \tag{2.9a}$$

If D is the domain of f, then

$$(x \in D) \Rightarrow (-x \in D) \tag{2.9b}$$

should hold.

■ **A:** $y = \cos x$, ■ **B:** $y = x^4 - 3x^2 + 1$.

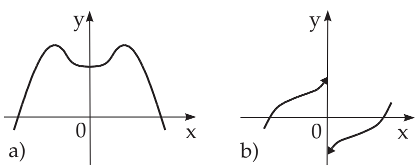


Figure 2.4

2.1.3.5 Odd Functions

Odd functions **(Fig. 2.4b)** satisfy the relation

$$f(-x) = -f(x). \tag{2.10a}$$

If D is the domain of f, then

$$(x \in D) \Rightarrow (-x \in D) \tag{2.10b}$$

should hold.

■ **A:** $y = \sin x$, ■ **B:** $y = x^3 - x$.

2.1.3.6 Representation with Even and Odd Functions

If for the domain D of a function f the condition "from $x \in D$ it follows that $-x \in D$" holds, then f can be written as a sum of an even function g and an odd function u:

$$f(x) = g(x) + u(x) \quad \text{with} \quad g(x) = \frac{1}{2}[f(x) + f(-x)], \quad u(x) = \frac{1}{2}[f(x) - f(-x)]. \tag{2.11}$$

■ $f(x) = e^x = \frac{1}{2}\left(e^x + e^{-x}\right) + \frac{1}{2}\left(e^x - e^{-x}\right) = \cosh x + \sinh x$ (see 2.9.1, p. 89).

2.1.3.7 Periodic Functions

Periodic functions satisfy the relation

$$f(x + T) = f(x), \quad T \text{ const}, \; T \neq 0. \tag{2.12}$$

Obviously, if the above equality holds for some T, it holds for any integer multiple of T. The smallest positive number T satisfying the relation is called the *period* **(Fig. 2.5)**.

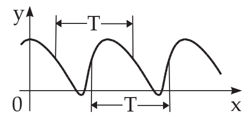


Figure 2.5

2.1.3.8 Inverse Functions

A function $y = f(x)$ with domain D and range W assigns a unique $y \in W$ to every $x \in D$. If reversed, to every $y \in W$ there belongs only one $x \in D$, then the *inverse function* of f can be defined. It is denoted by φ or by f^{-1}. Here f^{-1} is a symbol for a function, not a power of f.
To find the inverse function of f, the variables x and y are interchanged in the formula of f, then y is expressed from $x = f(x)$ in order to get $y = \varphi(x)$. The representations $y = f(x)$ and $x = \varphi(y)$ are equivalent. The following important formulas come from this relation

$$f(\varphi(y)) = y \quad \text{and} \quad \varphi(f(x)) = x. \tag{2.13}$$

The graph of an inverse function $y = \varphi(x)$ is obtained by reflection of the graph of $y = f(x)$ with respect to the line $y = x$ **(Fig. 2.6)**.

■ The function $y = f(x) = e^x$ $(D: -\infty < x < \infty, W: y > 0)$ is equivalent Obviously, every strictly monotonic function has an inverse function.

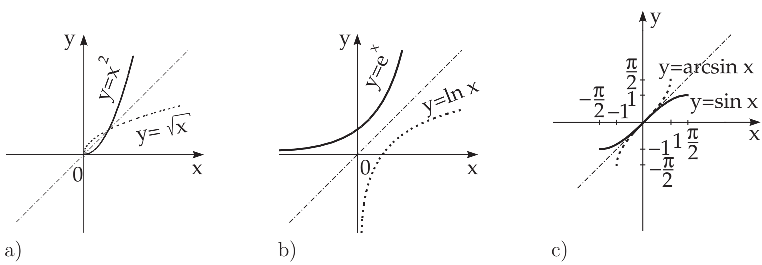


Figure 2.6

Examples of Inverse Functions:

■ **A:**	$y = f(x) = x^2$	with D:	$x \geq 0$,	W: $y \geq 0$;
	$y = \varphi(x) = \sqrt{x}$	with D:	$x \geq 0$,	W: $y \geq 0$.
■ **B:**	$y = f(x) = e^x$	with D:	$-\infty < x < \infty$,	W: $y > 0$;
	$y = \varphi(x) = \ln x$	with D:	$x > 0$,	W: $-\infty < y < \infty$.
■ **C:**	$y = f(x) = \sin x$	with D:	$-\pi/2 \leq x \leq \pi/2$,	W: $-1 \leq y \leq 1$;
	$y = \varphi(x) = \arcsin x$	with D:	$-1 \leq x \leq 1$,	W: $-\pi/2 \leq y \leq \pi/2$.

Remarks:
1. If a function f is strictly monotone in an interval $I \subset D$, then there is an inverse f^{-1} for this interval.
2. If a non-monotone function can be partitioned in strictly monotone parts, then the corresponding inverse exists for each part.

2.1.4 Limits of Functions

2.1.4.1 Definition of the Limit of a Function

The function $y = f(x)$ has the *limit* A at $x = a$

$$\lim_{x \to a} f(x) = A \quad \text{or} \quad f(x) \to A \quad \text{for} \quad x \to a, \tag{2.14}$$

if as x approaches the value a infinitely closely, the value of $f(x)$ approaches the value A infinitely closely. The function $f(x)$ does not have to be defined at a, and even if defined, it does not matter whether $f(a)$ is equal to A.

Precise Definition: The limit (2.14) exists, if for any given positive number ε there is a positive number η such that for every $x \neq a$ belonging to the domain and satisfying the inequality

$$|x - a| < \eta, \tag{2.15a}$$

the inequality

$$|f(x) - A| < \varepsilon \tag{2.15b}$$

holds eventually with the exception of the point a **(Fig. 2.7)**. If a is an endpoint of a connected region, then the inequality $|x - a| < \eta$ is reduced either to $a - \eta < x$ or to $x < a + \eta$ (see also 2.1.4.5).

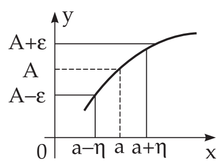


Figure 2.7

2.1.4.2 Definition by Limit of Sequences (see 7.1.2, p. 458)

A function $f(x)$ has the limit A at $x = a$ if for every sequence $x_1, x_2, \ldots, x_n, \ldots$ of the values of x from the domain and converging to a (but being not equal to a), the sequence of the corresponding values of the function $f(x_1), f(x_2), \ldots, f(x_n), \ldots$ converges to A.

2.1.4.3 Cauchy Condition for Convergence

A necessary and sufficient condition for a function $f(x)$ to have a limit at $x = a$ is that for any two values $x_1 \neq a$ and $x_2 \neq a$ belonging to the domain and being close enough to a, the values $f(x_1)$ and $f(x_2)$ are also close enough to each other.

Precise Definition: A necessary and sufficient condition for a function $f(x)$ to have a limit at $x = a$ is that for any given positive number ε there is a positive number η such that for arbitrary values x_1 and x_2 belonging to the domain and satisfying the inequalities

$$0 < |x_1 - a| < \eta \quad \text{and} \quad 0 < |x_2 - a| < \eta, \tag{2.16a}$$

the inequality

$$|f(x_1) - f(x_2)| < \varepsilon \tag{2.16b}$$

holds.

2.1.4.4 Infinity as a Limit of a Function

The symbol

$$\lim_{x \to a} |f(x)| = \infty \tag{2.17}$$

means that as x approaches a, the absolute value $|f(x)|$ does not have an upper bound.

Precise Definition: The equality (2.17) holds if for any given positive number K there is a positive number η such that for any $x \neq a$ from the interval

$$a - \eta < x < a + \eta \tag{2.18a}$$

the corresponding value of $|f(x)|$ is larger than K:

$$|f(x)| > K\,. \tag{2.18b}$$

If all the values of $f(x)$ in the interval

$$a - \eta < x < a + \eta \tag{2.18c}$$

are positive, one writes

$$\lim_{x \to a} f(x) = +\infty; \tag{2.18d}$$

if they are negative, one writes

$$\lim_{x \to a} f(x) = -\infty. \tag{2.18e}$$

2.1.4.5 Left-Hand and Right-Hand Limit of a Function

A function $f(x)$ has a *left-hand limit* A^- at $x = a$, if as x tends to a from the left, the value $f(x)$ tends to A^-:

$$A^- = \lim_{x \to a-0} f(x) = f(a - 0). \tag{2.19a}$$

Similarly, a function has a *right-hand limit* A^+ if as x tends to a from the right, the value $f(x)$ tends to A^+:

$$A^+ = \lim_{x \to a+0} f(x) = f(a + 0). \tag{2.19b}$$

The equality $\lim\limits_{x \to a} f(x) = A$ is valid only if the left-hand and right-hand limits exist, and they are equal:

$$A^+ = A^- = A. \tag{2.19c}$$

■ The function $f(x) = \dfrac{1}{1 + e^{\frac{1}{x-1}}}$ tends to different values from the left and from the right for $x \to 1$: $f(1-0) = 1,\ f(1+0) = 0$ **(Fig. 2.8)**.

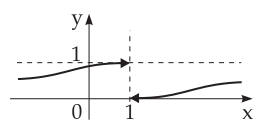


Figure 2.8

2.1.4.6 Limit of a Function as x Tends to Infinity

Case a) A number A is called the limit of a function $f(x)$ as $x \to +\infty$, and one writes

$$A = \lim_{x \to +\infty} f(x) \tag{2.20a}$$

if for any given positive number ε there is a number $N > 0$ such that for every $x > N$, the corresponding value $f(x)$ is in the interval $A - \varepsilon < f(x) < A + \varepsilon$. Analogously

$$A = \lim_{x \to -\infty} f(x) \tag{2.20b}$$

is the limit of a function $f(x)$ as $x \to -\infty$ if for any given positive number ε there is a positive number $N > 0$ such that for any $x < -N$ the corresponding value of $f(x)$ is in the interval $A-\varepsilon < f(x) < A+\varepsilon$.

■ **A:** $\lim\limits_{x \to +\infty} \dfrac{x+1}{x} = 1$, ■ **B:** $\lim\limits_{x \to -\infty} \dfrac{x+1}{x} = 1$, ■ **C:** $\lim\limits_{x \to -\infty} e^x = 0$.

Case b) Assume that for any positive number K, there is a positive number N such that if $x > N$ or $x < -N$ then the absolute value of the function is larger then K. In this case one writes

$$\lim_{x \to +\infty} |f(x)| = \infty \quad \text{or} \quad \lim_{x \to -\infty} |f(x)| = \infty. \tag{2.20c}$$

■ **A:** $\lim\limits_{x\to+\infty} \dfrac{x^3-1}{x^2} = +\infty$, ■ **B:** $\lim\limits_{x\to-\infty} \dfrac{x^3-1}{x^2} = -\infty$,

■ **C:** $\lim\limits_{x\to+\infty} \dfrac{1-x^3}{x^2} = -\infty$, ■ **D:** $\lim\limits_{x\to-\infty} \dfrac{1-x^3}{x^2} = +\infty$.

2.1.4.7 Theorems About Limits of Functions

1. Limit of a Constant Function The limit of a constant function is the constant itself:

$$\lim_{x\to a} A = A. \tag{2.21}$$

2. Limit of a Sum or a Difference If among a finite number of functions each has a limit, then the limit of their sum or difference is equal to the sum or difference of their limits (if this last expression does not contain $\infty - \infty$):

$$\lim_{x\to a} [f(x) + \varphi(x) - \psi(x)] = \lim_{x\to a} f(x) + \lim_{x\to a} \varphi(x) - \lim_{x\to a} \psi(x)\,. \tag{2.22}$$

3. Limit of Products If among a finite number of functions each has a limit, then the limit of their product is equal to the product of their limits (if this last expression does not contain a $0 \cdot \infty$ type):

$$\lim_{x\to a} [f(x)\,\varphi(x)\,\psi(x)] = \left[\lim_{x\to a} f(x)\right] \left[\lim_{x\to a} \varphi(x)\right] \left[\lim_{x\to a} \psi(x)\right]. \tag{2.23}$$

4. Limit of a Quotient The limit of the quotient of two functions is equal to the quotient of their limits, in the case when both limits exist and the limit of the denominator is not equal to zero (and this last expression is not an ∞/∞ type):

$$\lim_{x\to a} \frac{f(x)}{\varphi(x)} = \frac{\lim\limits_{x\to a} f(x)}{\lim\limits_{x\to a} \varphi(x)}. \tag{2.24}$$

Also if the denominator is equal to zero, usually one can tell if the limit exists or not, checking the sign of the denominator (the indeterminate form is $0/0$). Similarly, one can calculate the limit of a power by taking a suitable power of the limit (if it is not a 0^0, 1^∞, or ∞^0 type).

5. Pinching If the values of a function $f(x)$ lie between the values of the functions $\varphi(x)$ and $\psi(x)$, i.e., $\varphi(x) < f(x) < \psi(x)$, and if $\lim\limits_{x\to a} \varphi(x) = A$ and $\lim\limits_{x\to a} \psi(x) = A$ hold, then $f(x)$ has a limit, too, and

$$\lim_{x\to a} f(x) = A. \tag{2.25}$$

2.1.4.8 Calculation of Limits

The calculation of the value of a limit can be made by using the 5 described theorems as well as some transformations (see 2.1.4.7).

1. Suitable Transformations

For the calculation of limits the expression is to be transformed into a suitable form. There are several types of recommended transformations in different cases; here are three of them as examples.

■ **A:** $\lim\limits_{x\to 1} \dfrac{x^3-1}{x-1} = \lim\limits_{x\to 1} (x^2+x+1) = 3.$

■ **B:** $\lim\limits_{x\to 0} \dfrac{\sqrt{1+x}-1}{x} = \lim\limits_{x\to 0} \dfrac{(\sqrt{1+x}-1)(\sqrt{1+x}+1)}{x(\sqrt{1+x}+1)} = \lim\limits_{x\to 0} \dfrac{1}{\sqrt{1+x}+1} = \dfrac{1}{2}.$

■ **C:** $\lim\limits_{x\to 0} \dfrac{\sin 2x}{x} = \lim\limits_{x\to 0} \dfrac{2(\sin 2x)}{2x} = 2 \lim\limits_{2x\to 0} \dfrac{\sin 2x}{2x} = 2\,.$ Here one can refer to the well-known theorem

$$\lim_{\alpha\to 0}\frac{\sin\alpha}{\alpha}=1 \qquad \text{(see (■ A, 2.1.4.9, p. 57))}.$$

2. Bernoulli-l'Hospital Rule

In the case of indeterminate forms like $\frac{0}{0}$, $\frac{\infty}{\infty}$, $0\cdot\infty$, $\infty-\infty$, 0^0, ∞^0, 1^∞, one often applies the *Bernoulli-l'Hospital rule* (usually called *l'Hospital rule* for short):

Case a) Indeterminate Forms $\frac{\mathbf{0}}{\mathbf{0}}$ or $\frac{\infty}{\infty}$: First, use the theorem only after checking if for $f(x)=\frac{\varphi(x)}{\psi(x)}$ the following conditions are fulfilled.
Suppose $\lim\limits_{x\to a}\varphi(x)=0$ and $\lim\limits_{x\to a}\psi(x)=0$ or $\lim\limits_{x\to a}\varphi(x)=\infty$ and $\lim\limits_{x\to a}\psi(x)=\infty$, and suppose that there is an interval containing a such that the functions $\varphi(x)$ and $\psi(x)$ are defined and differentiable in this interval except perhaps at a, and $\psi'(x)\neq 0$ in this interval, and $\lim\limits_{x\to a}\frac{\varphi'(x)}{\psi'(x)}$ exists. Then

$$\lim_{x\to a}f(x)=\lim_{x\to a}\frac{\varphi(x)}{\psi(x)}=\lim_{x\to a}\frac{\varphi'(x)}{\psi'(x)}. \tag{2.26}$$

Remark: If the limit of the ratio of the derivatives does not exist, it does not mean that the original limit does not exist. Maybe it does, but one cannot tell this using l'Hospital's rule.
If $\lim\limits_{x\to a}\frac{\varphi'(x)}{\psi'(x)}$ is still an indeterminate form, and the numerator and denominator satisfy the assumptions of the above theorem, l'Hospital's rule can be used again.

■ $$\lim_{x\to 0}\frac{\ln\sin 2x}{\ln\sin x}=\lim_{x\to 0}\frac{\dfrac{2\cos 2x}{\sin 2x}}{\dfrac{\cos x}{\sin x}}=\lim_{x\to 0}\frac{2\tan x}{\tan 2x}=\lim_{x\to 0}\frac{\dfrac{2}{\cos^2 x}}{\dfrac{2}{\cos^2 2x}}=\lim_{x\to 0}\frac{\cos^2 2x}{\cos^2 x}=1.$$

Case b) Indeterminate Form $\mathbf{0}\cdot\infty$: Having $f(x)=\varphi(x)\,\psi(x)$ and $\lim\limits_{x\to a}\varphi(x)=0$ and $\lim\limits_{x\to a}\psi(x)=\infty$, then in order to use l'Hospital's rule for $\lim\limits_{x\to a}f(x)$ it is to be transformed into one of the forms $\lim\limits_{x\to a}\frac{\varphi(x)}{\frac{1}{\psi(x)}}$ or $\lim\limits_{x\to a}\frac{\psi(x)}{\frac{1}{\varphi(x)}}$, so it is reduced to an indeterminate form $\frac{0}{0}$ or $\frac{\infty}{\infty}$ like in case **a)**.

■ $$\lim_{x\to\pi/2}(\pi-2x)\tan x=\lim_{x\to\pi/2}\frac{\pi-2x}{\cot x}=\lim_{x\to\pi/2}\frac{-2}{-\dfrac{1}{\sin^2 x}}=2.$$

Case c) Indeterminate Form $\infty-\infty$: If $f(x)=\varphi(x)-\psi(x)$ and $\lim\limits_{x\to a}\varphi(x)=\infty$ and $\lim\limits_{x\to a}\psi(x)=\infty$ hold, then this expression can be transformed into the form $\frac{0}{0}$ or $\frac{\infty}{\infty}$ usually in several different ways; for instance as $\varphi-\psi=\left(\frac{1}{\psi}-\frac{1}{\varphi}\right)\Big/\frac{1}{\varphi\psi}$. Then it is to proceed as in case **a)**.

■ $\lim\limits_{x\to 1}\left(\frac{x}{x-1}-\frac{1}{\ln x}\right)=\lim\limits_{x\to 1}\left(\frac{x\ln x-x+1}{x\ln x-\ln x}\right)=\frac{0}{0}$. Applying l'Hospital rule twice yields

$$\lim_{x\to 1}\left(\frac{x\ln x-x+1}{x\ln x-\ln x}\right)=\lim_{x\to 1}\left(\frac{\ln x}{\ln x+1-\dfrac{1}{x}}\right)=\lim_{x\to 1}\left(\frac{\dfrac{1}{x}}{\dfrac{1}{x}+\dfrac{1}{x^2}}\right)=\frac{1}{2}.$$

Case d) Indeterminate Forms $\mathbf{0^0}$, $\infty^{\mathbf{0}}$, $\mathbf{1}^\infty$: If $f(x)=\varphi(x)^{\psi(x)}$ and $\lim\limits_{x\to a}\varphi(x)=+0$ and $\lim\limits_{x\to a}\psi(x)=$

0 holds, then first the limit A of $\ln f(x) = \psi(x)\ln\varphi(x)$, is to be calculated, which has the form $0\cdot\infty$ (case **b)**), then $\lim\limits_{x\to a} f(x) = e^A$ holds.

The procedures in the cases ∞^0 and 1^∞ are similar.

■ $\lim\limits_{x\to+0} x^x = X$, $\ln x^x = x\ln x$, $\lim\limits_{x\to+0} x\ln x = \lim\limits_{x\to+0}\dfrac{\ln x}{x^{-1}} = \lim\limits_{x\to+0}(-x) = 0$, i.e., $A = \ln X = 0$, so $X = 1$, and finally $\lim\limits_{x\to+0} x^x = 1$.

3. Taylor Expansion

Besides l'Hospital's rule the expansion of functions of indeterminate form into Taylor series can be applied (see 6.1.4.5, p. 442).

■ $$\lim_{x\to 0}\frac{x-\sin x}{x^3} = \lim_{x\to 0}\frac{x-\left(x-\dfrac{x^3}{3!}+\dfrac{x^5}{5!}-\cdots\right)}{x^3} = \lim_{x\to 0}\left(\frac{1}{3!}-\frac{x^2}{5!}+\cdots\right) = \frac{1}{6}.$$

2.1.4.9 Order of Magnitude of Functions and Landau Order Symbols

Comparing two functions, often their mutual behavior with respect to a certain argument $x = a$ is to be considered. It is also convenient to compare the order of magnitude of the functions.

1. A function $f(x)$ tends to infinity at a at a higher order (or faster rate) than a function $g(x)$ at a, if the quotient $\left|\dfrac{f(x)}{g(x)}\right|$ and the absolute values of $f(x)$ exceed any limit as x tends to a.

2. A function $f(x)$ tends to zero at a at a higher order than a function $g(x)$ at a, if the absolute values of $f(x), g(x)$ and the quotient $\dfrac{f(x)}{g(x)}$ tends to zero as x tends to a.

3. Two functions $f(x)$ and $g(x)$ tend to zero or to infinity at s at the same order of magnitude, if $0 < m < \left|\dfrac{f(x)}{g(x)}\right| < M$ holds for the absolute value of their quotient as x tends to a, where M and m are constants.

4. Landau Order Symbols The mutual behavior of two functions at a point $x = a$ can be described by the *Landau order symbols* O ("big O"), or o ("small o") as follows: If $x \to a$ then

$$f(x) = O(g(x)) \quad \text{means that} \quad \lim_{x\to a}\frac{f(x)}{g(x)} = A \neq 0, \quad A = \text{const}, \tag{2.27a}$$

and

$$f(x) = o(g(x)) \quad \text{means that} \quad \lim_{x\to a}\frac{f(x)}{g(x)} = 0, \tag{2.27b}$$

where $a = \pm\infty$ is also possible. The Landau order symbols have meaning only by assuming x tends to a given a.

■ **A:** $\sin x = O(x)$ for $x \to 0$, because with $f(x) = \sin x$ and $g(x) = x$ holds: $\lim\limits_{x\to 0}\dfrac{\sin x}{x} = 1 \neq 0$, i.e., $\sin x$ behaves like x in the neighborhood of $x = 0$.

■ **B:** For $f(x) = 1-\cos x$ and $g(x) = \sin x$ the function $f(x)$ vanishes with a higher order than $g(x)$: $\lim\limits_{x\to 0}\left|\dfrac{f(x)}{g(x)}\right| = \lim\limits_{x\to 0}\left|\dfrac{1-\cos x}{\sin x}\right| = 0$, i.e., $1-\cos x = o(\sin x)$ for $x \to 0$.

■ **C:** $f(x)$ and $g(x)$ vanish by the same order for $f(x) = 1-\cos x$, $g(x) = x^2$:

$$\lim_{x\to 0}\left|\frac{f(x)}{g(x)}\right| = \lim_{x\to 0}\left|\frac{1-\cos x}{x^2}\right| = \frac{1}{2}\,,\ \text{i.e.,}\ 1-\cos x = O(x^2)\ \text{for}\ x\to 0.$$

5. Polynomial The order of magnitude of polynomials at $\pm\infty$ can be expressed by their degree. So the function $f(x) = x$ has order 1, a polynomial of degree $n+1$ has an order higher by one than a polynomial of degree n.

6. Exponential Function The exponential function e^x tends faster to infinity for $x \to \infty$ more quickly to infinity than any high power x^n (n is a fixed positive number):

$$\lim_{x\to\infty}\left|\frac{e^x}{x^n}\right| = \infty. \tag{2.28a}$$

The proof follows by applying l'Hospital's rule for a natural number n:

$$\lim_{x\to\infty}\frac{e^x}{x^n} = \lim_{x\to\infty}\frac{e^x}{nx^{n-1}} = \ldots = \lim_{x\to\infty}\frac{e^x}{n!} = \infty. \tag{2.28b}$$

7. Logarithmic Function The logarithm tends to infinity more slowly than any small positive power x^α (α is a fixed positive number):

$$\lim_{x\to\infty}\left|\frac{\log x}{x^\alpha}\right| = 0. \tag{2.29}$$

The proof is with the help of l'Hospital's rule.

2.1.5 Continuity of a Function

2.1.5.1 Notion of Continuity and Discontinuity

Most functions occurring in practice are continuous, i.e., for small changes of the argument x a *continuous function* $y(x)$ changes also only a little. The graphical representation of such a function results in a continuous curve. If the curve is broken at some points, the corresponding function is *discontinuous*, and the values of the arguments where the breaks are, are the *points of discontinuity*. **Fig. 2.9** shows the curve of a function, which is *piecewise continuous*. The points of discontinuity are A, B, C, D, E, F and G. The arrow-heads show that the endpoints do not belong to the curve.

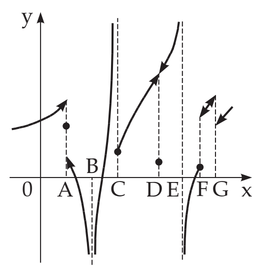


Figure 2.9

2.1.5.2 Definition of Continuity

A function $y = f(x)$ is called *continuous* at the point $x = a$ if

1. $f(x)$ is defined at a;
2. the limit $\lim\limits_{x\to a} f(x)$ exists and is equal to $f(a)$.

This is exactly the case if for an arbitrary $\varepsilon > 0$ there is a $\delta(\varepsilon) > 0$ such that

$$|f(x) - f(a)| < \varepsilon \quad \text{for every } x \text{ with} \quad |x-a| < \delta \tag{2.30}$$

holds.

Here also it is to talk about *one-sided* (*left-* or *right-hand sided*) *continuity*, if instead of $\lim\limits_{x\to a} f(x) = f(a)$ only the one-sided limit $\lim\limits_{x\to a-0} f(x)$ (or $\lim\limits_{x\to a+0} f(x)$) is to be considered and this is equal to the value $f(a)$.

If a function is continuous for every x in a given interval from a to b, then the function is called *continuous in this interval*, which can be open, half-open, or closed (see 1.1.1.3, **3.**, p. 2). If a function is defined and continuous at every point of the numerical axis, it is called *continuous everywhere*.

A function has a *point of discontinuity* at $x = a$, which is an interior point or an endpoint of its domain,

if the function is not defined here, or $f(a)$ is not equal to the limit $\lim_{x\to a} f(x)$, or the limit does not exist. If the function is defined only on one side of $x = a$, e.g., $+\sqrt{x}$ for $x = 0$ and $\arccos x$ for $x = 1$, then it is not a point of discontinuity but it is a termination.
A function $f(x)$ is called *piecewise continuous*, if it is continuous at every point of an interval except at a finite number of points, and at these points it has finite jumps.

2.1.5.3 Most Frequent Types of Discontinuities

1. Values of the Function Tend to Infinity

The most frequent discontinuity occurs if the function tends to $\pm\infty$ (points B, C, and E in **Fig. 2.9**).

■ **A:** $f(x) = \tan x,\ f\left(\frac{\pi}{2} - 0\right) = +\infty,\ f\left(\frac{\pi}{2} + 0\right) = -\infty$. The type of discontinuity (see **Fig. 2.34**, p. 78) is the same as at E in **Fig. 2.9**. For the meaning of the symbols $f(a-0),\ f(a+0)$ see 2.1.4.5, p. 54.

■ **B:** $f(x) = \frac{1}{(x-1)^2}$, $f(1-0) = +\infty,\ f(1+0) = +\infty$. The type of discontinuity is the same as at the point B in **Fig. 2.9**.

■ **C:** $f(x) = e^{\frac{1}{x-1}}$, $f(1-0) = 0,\ f(1+0) = \infty$. The type of discontinuity is the same as at C in **Fig. 2.9**, with the difference that this function $f(x)$ is not defined at $x = 1$.

2. Finite Jump

Passing through $x = a$ the function $f(x)$ jumps from a finite value to another finite value (like at the points A, F, G in **Fig. 2.9**, p. 58): The value of the function $f(x)$ for $x = a$ may not be defined here, as at point G; or it can coincide with $f(a-0)$ or with $f(a+0)$ (point F); or it can be different from $f(a-0)$ and $f(a+0)$ (point A).

■ **A:** $f(x) = \frac{1}{1 + e^{\frac{1}{x-1}}}$, $f(1-0) = 1,\ f(1+0) = 0$ (**Fig. 2.8**, p. 54).

■ **B:** $f(x) = E(x)$ (**Fig. 2.1c**, p. 50) $f(n-0) = n-1,\ f(n+0) = n$ (n integer).

■ **C:** $f(x) = \lim_{n\to\infty} \frac{1}{1 + x^{2n}}$, $f(1-0) = 1,\ f(1+0) = 0,\ f(1) = \frac{1}{2}$.

3. Removable Discontinuity

Assuming that $\lim_{x\to a} f(x)$ exists, i.e., $f(a-0) = f(a+0)$ holds, but either the function is not defined for $x = a$ or there is $f(a) \neq \lim_{x\to a} f(x)$ (point D in **Fig. 2.9**, p. 58). This type of discontinuity is called *removable*, because defining $f(a) = \lim_{x\to a} f(x)$ the function becomes continuous here. The procedure consists of adding only one point to the curve, or changing the place only of one point at D. The different indeterminate expressions for $x = a$, which have a finite limit examined by l'Hospital's rule or with other methods, are examples of removable discontinuities.

■ $f(x) = \frac{\sqrt{1+x} - 1}{x}$ is an undetermined $\frac{0}{0}$ expression for $x = 0$, but $\lim_{x\to 0} f(x) = \frac{1}{2}$; the function

$$f(x) = \begin{cases} \dfrac{\sqrt{1+x} - 1}{x} & \text{for } x \neq 0 \\[2ex] \dfrac{1}{2} & \text{for } x = 0 \end{cases}$$

is continuous.

2.1.5.4 Continuity and Discontinuity of Elementary Functions

The elementary functions are continuous on their domains; the points of discontinuity do not belong to their domain. The following theorems hold:

1. Polynomials are continuous everywhere.

2. Rational Functions $\frac{\boldsymbol{P(x)}}{\boldsymbol{Q(x)}}$ with polynomials $P(x)$ and $Q(x)$ are continuous everywhere except the points x, where $Q(x) = 0$. If at $x = a$, $Q(a) = 0$ and $P(a) \neq 0$, the function tends to $\pm\infty$ on both sides of a; this point is called a *pole*. The function also has a pole if $P(a) = 0$, but a is a root of the denominator with higher multiplicity than for the numerator (see 1.6.3.1, **2.,** p. 43). Otherwise the discontinuity is removable.

3. Irrational Functions Roots of polynomials are continuous for every x in their domain. At the end of the domain they can terminate by a finite value if the radicand changes its sign. Roots of rational functions are discontinuous for such values of x where the radicand is discontinuous.

4. Trigonometric Functions The functions $\sin x$ and $\cos x$ are continuous everywhere; $\tan x$ and $\sec x$ have infinite jumps at the points $x = \frac{(2n+1)\pi}{2}$; the functions $\cot x$ and $\operatorname{cosec} x$ have infinite jumps at the points $x = n\pi$ (n integer).

5. Inverse Trigonometric Functions The functions $\arctan x$ and $\operatorname{arccot} x$ are continuous everywhere, $\arcsin x$ and $\arccos x$ terminate at the end of their domain because of $-1 \leq x \leq +1$, and they are continuous here from one side.

6. Exponential Functions $\boldsymbol{e^x}$ or $\boldsymbol{a^x}$ with $\boldsymbol{a > 0}$ They are continuous everywhere.

7. Logarithmic Function $\boldsymbol{\log x}$ with Arbitrary Positive Base The function is continuous for all positive x and terminates at $x = 0$ because of $\lim\limits_{x\to+0} \log x = -\infty$ by a right-sided limit.

8. Composite Elementary Functions The continuity is to be checked for every point x of every elementary function containing in the composition (see also continuity of composite functions in 2.1.5.5, **2.**, p. 61).

■ Find the points of discontinuity of the function $y = \frac{e^{\frac{1}{x-2}}}{x \sin \sqrt[3]{1-x}}$. The exponent $\frac{1}{x-2}$ has an infinite jump at $x = 2$; for $x = 2$ also $e^{\frac{1}{x-2}}$ has an infinite jump: $\left(e^{\frac{1}{x-2}}\right)_{x=2-0} = 0$, $\left(e^{\frac{1}{x-2}}\right)_{x=2+0} = \infty$. The function y has a finite denominator at $x = 2$. Consequently, at $x = 2$ there is an infinite jump of the same type as at point C in **Fig. 2.9**, p. 58.
For $x = 0$ the denominator is also zero, just like for the values of x, for which $\sin \sqrt[3]{1-x}$ is equal to zero. These last ones correspond to the roots of the equation $\sqrt[3]{1-x} = n\pi$ or $x = 1 - n^3\pi^3$, where n is an arbitrary integer. The numerator is not equal to zero for these numbers, so at the points $x = 0$, $x = 1$, $x = 1 \pm \pi^3$, $x = 1 \pm 8\pi^3$, $x = 1 \pm 27\pi^3, \ldots$ the function has the same type of discontinuity as the point E in **Fig. 2.9**, p. 58.

2.1.5.5 Properties of Continuous Functions

1. Continuity of Sum, Difference, Product and Quotient of Continuous Functions

If $f(x)$ and $g(x)$ are continuous on the interval $[a, b]$, then $f(x) \pm g(x)$, $f(x)\,g(x)$ are also continuous, and if $g(x) \neq 0$ on this interval, then $\frac{f(x)}{g(x)}$ is also continuous.

2. Continuity of Composite Functions $y = f(u(x))$

If $u(x)$ is continuous at $x = a$ and $f(u)$ is continuous at $u = u(a)$ then the composite function $y = f(u(x))$ is continuous at $x = a$, and

$$\lim_{x \to a} f(u(x)) = f\left(\lim_{x \to a} u(x)\right) = f(u(a)) \tag{2.31}$$

is valid. This means that a continuous function of a continuous function is also continuous.

Remark: The converse sentence is not valid. It is possible that the composite function of discontinuous functions is continuous.

3. Bolzano Theorem

If a function $f(x)$ is continuous on a finite closed interval $[a, b]$, and $f(a)$ and $f(b)$ have different signs, then $f(x)$ has at least one root in this interval, i.e., there exists at least one interior point of this interval c such that:

$$f(c) = 0 \quad \text{with} \quad a < c < b. \tag{2.32}$$

The geometric interpretation of this statement is that the graph of a continuous function can go from one side of the x-axis to the other side only if the curve has an intersection point with the x-axis.

4. Intermediate Value Theorem

If a function $f(x)$ is continuous on an interval, and it has different values A and B, at the points a and b of this interval, where $a < b$, i.e.,

$$f(a) = A, \quad f(b) = B, \quad A \neq B, \tag{2.33a}$$

then for any value C between A and B there is at least one point c between a and b such that

$$f(c) = C, \quad (a < c < b, \quad A < C < B \text{ or } A > C > B). \tag{2.33b}$$

In other words: The function $f(x)$ takes every value between A and B on the interval (a, b) at least once. Or: The continuous image of an interval is an interval.

5. Existence of an Inverse Function

If a one-to-one function is continuous on an interval, it is strictly monotone on this interval.
If a function $f(x)$ is continuous on a connected domain I, and it is strictly monotone increasing or decreasing, then for this $f(x)$ there also exists a continuous, strictly monotone increasing or decreasing inverse function $\varphi(x)$ (see also 2.1.3.8, p. 52), which is defined on domain II given by the values of $f(x)$ (**Fig. 2.10**).

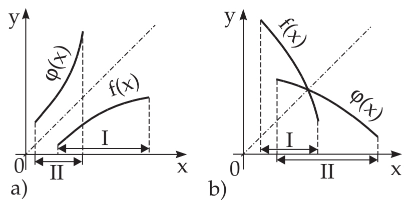


Figure 2.10

Remark: In order to make sure that the inverse function of $f(x)$ is continuous, $f(x)$ must be continuous on an interval. Supposing only that the function is strictly monotonic on an interval, and continuous at an interior point c, and $f(c) = C$, then the inverse function exists, but may be not continuous at C.

6. Theorem About the Boundedness of a Function

If a function $f(x)$ is continuous on a finite, closed interval $[a, b]$ then it is bounded on this interval, i.e., there exist two numbers m and M such that

$$m \leq f(x) \leq M \quad \text{for} \quad a \leq x \leq b. \tag{2.34}$$

7. Weierstrass Theorem

If the function $f(x)$ is continuous on the finite, closed interval $[a, b]$ then $f(x)$ has an *absolute maximum* M and an *absolute minimum* m, i.e., there exists in this interval at least one point c and at least one point d such that for all x with $a \leq x \leq b$:

$$m = f(d) \leq f(x) \leq f(c) = M. \tag{2.35}$$

The difference between the greatest and smallest value of a continuous function is called its *variation* in the given interval. The notion of variation can be extended to the case when the function does not have any greatest or smallest value.

2.2 Elementary Functions

Elementary functions are defined by formulas containing a finite number of operations on the independent variable and constants. The operations are the four basic arithmetical operations, taking powers and roots, the use of an exponential or a logarithm function, or the use of trigonometric functions or inverse trigonometric functions. To distinguish are *algebraic* and *transcendental* elementary functions. As another type of function, can be defined the *non-elementary functions* (see for instance 8.2.5, p. 513).

2.2.1 Algebraic Functions

In an *algebraic function* the argument x and the function y are connected by an *algebraic equation*. It has the form

$$p_0(x) + p_1(x)y + p_2(x)y^2 + \ldots + p_n(x)y^n = 0 \tag{2.36}$$

where p_0, p_1,..., p_n are polynomials in x.

■ $3xy^3 - 4xy + x^3 - 1 = 0$, i.e., $p_0(x) = x^3 - 1$, $p_1(x) = -4x$, $p_2(x) = 0$, $p_3(x) = 3x$.

If it is possible to solve an algebraic equation (2.36) for y, then there is one of the following types of the simplest algebraic functions.

2.2.1.1 Polynomials

Performing only addition, subtraction and multiplication on the argument x then:

$$y = a_n x^n + a_{n-1}x^{n-1} + \ldots + a_0. \tag{2.37}$$

holds. In particular one can distinguish $y = a$ as a *constant*, $y = ax + b$ as a *linear function*, and $y = ax^2 + bx + c$ as a *quadratic function*.

2.2.1.2 Rational Functions

A rational function can always be written in the form of the ratio of two polynomials:

$$y = \frac{a_n x^n + a_{n-1}x^{n-1} + \ldots + a_0}{b_m x^m + b_{m-1}x^{m-1} + \ldots + b_0}. \tag{2.38a}$$

The special case

$$y = \frac{ax + b}{cx + d} \tag{2.38b}$$

is called a *homographic or linear fractional function*.

2.2.1.3 Irrational Functions

Besides the operations enumerated for rational functions, the argument x also occurs under the radical sign.

■ **A:** $y = \sqrt{2x + 3}$, ■ **B:** $y = \sqrt[3]{(x^2 - 1)\sqrt{x}}$.

2.2.2 Transcendental Functions

Transcendental functions cannot be given by an algebraic equation like (2.36). In the following paragraphs the simplest elementary transcendental functions are introduced.

2.2.2.1 Exponential Functions

The variable x or an algebraic function of x is in the exponent of a constant base (see 2.6.1, p. 72).

■ **A:** $y = e^x$, ■ **B:** $y = a^x$, ■ **C:** $y = 2^{3x^2-5x}$.

2.2.2.2 Logarithmic Functions

The function is the logarithm with a constant base of the variable x or an algebraic function of x (see 2.6.2, p. 73).

■ **A:** $y = \ln x\,$, ■ **B:** $y = \lg x\,$, ■ **C:** $y = \log_2(5x^2 - 3x)\,$.

2.2.2.3 Trigonometric Functions

The variable x or an algebraic function of x occurs under the symbols sin, cos, tan, cot, sec, cosec (see 2.7, p. 76).

■ **A:** $y = \sin x$, ■ **B:** $y = \cos(2x + 3)$, ■ **C:** $y = \tan\sqrt{x}$.

In general, the argument of a trigonometric function is not only an angle or a circular arc as in the geometric definition, but an arbitrary quantity. The trigonometric functions can be defined in a purely analytic way without any geometry. For instance one can represent them by an expansion in a series, or, e.g., the sin function as the solution of the differential equation $\frac{d^2y}{dx^2} + y = 0$ with the initial values $y = 0$ and $\frac{dy}{dx} = 1$ at $x = 0$. The numerical value of the argument of the trigonometric function is equal to the *arc* in units of radians. When dealing with trigonometric functions, the argument is considered to be given in *radian measure* (see 3.1.1.5, p. 131).

2.2.2.4 Inverse Trigonometric Functions

The variable x or an algebraic function of x is in the argument of the inverse trigonometric functions (see 2.8, p. 85) arcsin, arccos, etc.

■ **A:** $y = \arcsin x$, ■ **B:** $y = \arccos\sqrt{1 - x}\,$.

2.2.2.5 Hyperbolic Functions

(see 2.9, p. 89).

2.2.2.6 Inverse Hyperbolic Functions

(see 2.10, p. 93).

2.2.3 Composite Functions

Composite functions are all possible compositions of the above algebraic and transcendental functions, i.e., if a function has another function as an argument.

■ **A:** $y = \ln\sin x$, ■ **B:** $y = \dfrac{\ln x + \sqrt{\arcsin x}}{x^2 + 5e^x}$.

Such composition of a finite number of elementary functions again yields an elementary function. The examples **C** in the previous types of functions are also composite functions.

2.3 Polynomials

2.3.1 Linear Function

The graph of the *linear function*

$$y = ax + b \tag{2.39}$$

(polynomial of degree 1) is a *line* **(Fig. 2.11a)**. The *proportional factor* is denoted by a, the crossing point of the line and the axis of the ordinate by b.

For $a > 0$ the function is monotone increasing, for $a < 0$ it is monotone decreasing; for $a = 0$ it is a

polynomial of degree zero, i.e., it is a constant function. The intercepts are at $A\left(-\frac{b}{a}, 0\right)$ and $B(0, b)$ (for details see 3.5.2.6, **1.**, p. 195). With $b = 0$ *direct proportionality*

$$y = ax; \tag{2.40}$$

holds, graphically it is a line running through the origin **(Fig. 2.11b)**.

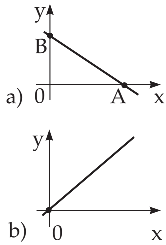


Figure 2.11

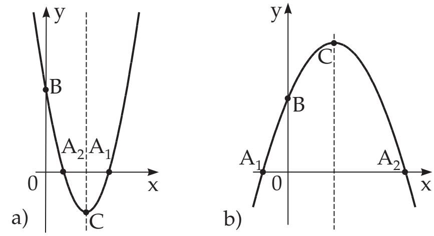


Figure 2.12

2.3.2 Quadratic Polynomial

The *polynomial of second degree*

$$y = ax^2 + bx + c \tag{2.41}$$

(quadratic polynomial) defines a *parabola* with a vertical axis of symmetry at $x = -\frac{b}{2a}$ **(Fig. 2.12)**. For $a > 0$ the function is first monotone decreasing, it has a minimum, then it is monotone increasing. For $a < 0$ first it is monotone increasing, it has a maximum, then it is monotone decreasing. In the case $b^2 - 4ac > 0$: The intersection points A_1, A_2 with the x-axis are $\left(\frac{-b \pm \sqrt{b^2 - 4ac}}{2a}, 0\right)$, the intersection point B with the y-axis, is at $(0, c)$. In the case $b^2 - 4ac = 0$ there is one intersection point (contact point) with the x-axis. In the case $b^2 - 4ac < 0$ there is no intersection point. The extremum point of the curve is at $C\left(-\frac{b}{2a}, \frac{4ac - b^2}{4a}\right)$ (for more details about the parabola see 3.5.2.10, p. 204).

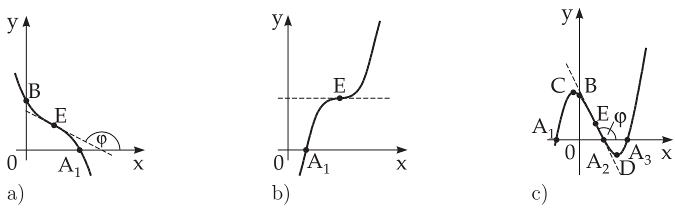


Figure 2.13

2.3.3 Cubic Polynomials

The *polynomial of third degree*

$$y = ax^3 + bx^2 + cx + d \tag{2.42}$$

defines a *cubic parabola* **(Fig. 2.13a,b,c)**. Both the shape of the curve and the behavior of the function depend on a and the discriminant $\Delta = 3ac - b^2$. If $\Delta \geq 0$ holds **(Fig. 2.13a,b)**, then for $a > 0$ the function is monotonically increasing, and for $a < 0$ it is decreasing. If $\Delta < 0$ the function has exactly one local minimum and one local maximum **(Fig. 2.13c)**. For $a > 0$ the value of the function rises from $-\infty$ until the maximum, then falls until the minimum, then it rises again to $+\infty$; for $a < 0$ the value of the function falls from $+\infty$ until the minimum, then rises until the maximum, then it falls again to $-\infty$. The intersection points with the x-axis are at the values of the real roots of (2.42) for $y = 0$. The function can have one, two (then there is a point where the x-axis is the tangent line of the curve) or three real roots: A_1, A_2 and A_3. The intersection point with the y-axis is at $B(0, d)$, the extreme points of the curve C and D, if any, are at $\left(-\frac{b \pm \sqrt{-\Delta}}{3a}, \frac{d + 2b^3 - 9abc \pm (6ac - 2b^2)\sqrt{-\Delta}}{27a^2}\right)$.

The inflection point which is also the center of symmetry of the curve is at $E\left(-\frac{b}{3a}, \frac{2b^3 - 9abc}{27a^2} + d\right)$.

At this point the tangent line has the slope $\tan\varphi = \left(\frac{dy}{dx}\right)_E = \frac{\Delta}{3a}$.

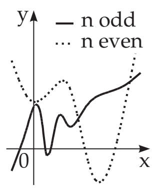


Figure 2.14

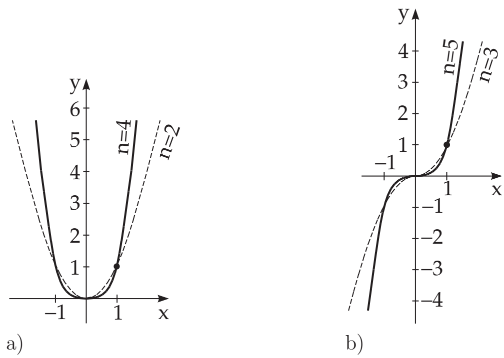


Figure 2.15

2.3.4 Polynomials of n-th Degree

The *integral rational function of n-th degree*

$$y = a_n x^n + a_{n-1}x^{n-1} + \ldots + a_1 x + a_0 \tag{2.43}$$

defines a *curve of n-th degree* or *n-th order* (see 3.5.2.5, p. 195) of *parabolic type* **(Fig. 2.14)**.

Case 1, n odd: For $a_n > 0$ the value of y changes continuously from $-\infty$ to $+\infty$, and for $a_n < 0$ from $+\infty$ to $-\infty$. The curve can intersect or contact the x-axis up to n times, and there is at least one intersection point (for the solution of an equation of n-th degree see 1.6.3.1, p. 43 and 19.1.2, p. 952). The function (2.43) has none or an even number up to $n - 1$ of extreme values, where minima and maxima occur alternately; the number of inflection points is odd and is between 1 and $n - 2$. There are no asymptotes or singularities.

Case 2, n even: For $a_n > 0$ the value of y changes continuously from $+\infty$ through its minimum until $+\infty$ and for $a_n < 0$ from $-\infty$ through its maximum until $-\infty$. The curve can intersect or contact the x-axis up to n times, but it is also possible that it never does that. The number of extrema is odd, and maxima and minima alternate; the number of inflection points is even, and it can also be zero. There are no asymptotes or singularities.

Before sketching the graph of a function, it is recommended first to determine the extreme points, the inflection points, the values of the first derivative at these points, then to sketch the tangent lines at these points, and finally to connect these points continuously.

2.3.5 Parabola of n-th Degree

The graph of the function

$$y = ax^n \tag{2.44}$$

where $n > 0$, integer, is a *parabola of n-th degree, or of n-th order* **(Fig. 2.15)**.

1. Special Case $a = 1$: The curve $y = x^n$ goes through the point $(0, 0)$ and $(1, 1)$ and contacts or intersects the x-axis at the origin. For even n the curve is symmetric with respect to the y-axis, and with a minimum at the origin. For odd n the curve is symmetric with respect to the origin, and it has an inflection point there. There is no asymptote.

2. General Case $a \neq 0$: The curve of $y = ax^n$ can be got from the curve of $y = x^n$ by stretching the ordinates by the factor $|a|$. For $a < 0$ the curve $y = |a|x^n$ is to be reflected with respect to the x-axis.

2.4 Rational Functions

2.4.1 Special Fractional Linear Function (Inverse Proportionality)

The graph of the function

$$y = \frac{a}{x} \tag{2.45}$$

is an *equilateral hyperbola*, whose asymptotes are the coordinate axes **(Fig. 2.16)**. The point of discontinuity is at $x = 0$ with $y = \pm\infty$. If $a > 0$ holds, then the function is strictly monotone decreasing in the interval $(-\infty, 0)$ with values from 0 to $-\infty$ and also strictly monotone decreasing in the interval $(0, +\infty)$ with values from $+\infty$ to 0 (curve in the first and third quadrants). If $a < 0$, then the function is increasing in the interval $(-\infty\,, 0)$ with values from 0 to $+\infty$ and also increasing in the interval $(0, +\infty)$ with values from $-\infty$ to 0 (dotted curve in the second and fourth quadrants). The vertices A and B are at $\left(\pm\sqrt{|a|}, +\sqrt{|a|}\right)$ and $\left(\pm\sqrt{|a|}, -\sqrt{|a|}\right)$ with the same sign for $a > 0$ and with different sign for $a < 0$. There are no extreme values (for more details about hyperbolas see 3.5.2.9, p. 201).

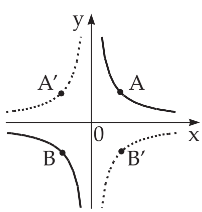


Figure 2.16

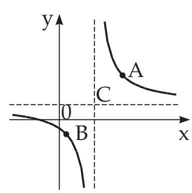


Figure 2.17

2.4.2 Linear Fractional Function

The graph of the function

$$y = \frac{a_1 x + b_1}{a_2 x + b_2} \qquad \left(a_2 \neq 0\,,\ \Delta = \begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix} = a_1 b_2 - b_1 a_2 \neq 0\right) \tag{2.46}$$

is an *equilateral hyperbola*, whose asymptotes are parallel to the coordinate axes **(Fig. 2.17)**. The center is at $C\left(-\frac{b_2}{a_2}, \frac{a_1}{a_2}\right)$. The parameter a in the equality (2.45) corresponds here to $-\frac{\Delta}{{a_2}^2}$ with $\Delta = \begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix}$. The vertices of the hyperbola $A \neq 0$ and B are at $\left(-\frac{b_2 \pm \sqrt{|\Delta|}}{a_2}, \frac{a_1 + \sqrt{|\Delta|}}{a_2}\right)$ and $\left(-\frac{b_2 \pm \sqrt{|\Delta|}}{a_2}, \frac{a_1 - \sqrt{|\Delta|}}{a_2}\right)$, where for $\Delta < 0$ the same signs are taken, for $\Delta > 0$ the different ones.

The point of discontinuity is at $x = -\frac{b_2}{a_2}$. For $\Delta < 0$ the values of the function are decreasing from $\frac{a_1}{a_2}$ to $-\infty$ and from $+\infty$ to $\frac{a_1}{a_2}$. For $\Delta > 0$ the values of the function are increasing from $\frac{a_1}{a_2}$ to $+\infty$ and from $-\infty$ to $\frac{a_1}{a_2}$. There is no extremum.

2.4.3 Curves of Third Degree, Type I

The graph of the function

$$y = a + \frac{b}{x} + \frac{c}{x^2} \quad \left(= \frac{ax^2 + bx + c}{x^2}\right) \quad (b \neq 0,\ \ c \neq 0) \tag{2.47}$$

(Fig. 2.18) is a *curve of third degree* (type I). It has two asymptotes $x = 0$ and $y = a$ and it has two branches. One of them corresponds to the monotone changing of y while it takes its values between a and $+\infty$ or $-\infty$; the other branch goes through three characteristic points: the intersection point with the asymptote $y = a$ at $A\left(-\frac{c}{b}, a\right)$, an extreme point at $B\left(-\frac{2c}{b}, a - \frac{b^2}{4c}\right)$ and an inflection point at $C\left(-\frac{3c}{b}, a - \frac{2b^2}{9c}\right)$. The positions of the branches depend on the signs of b and c, and there are four cases **(Fig. 2.18)**. The intersection points D, E with the x-axis, if any, are for $a \neq 0$ at $\left(\frac{-b \pm \sqrt{b^2 - 4ac}}{2a}, 0\right)$, for $a = 0$ at $\left(-\frac{c}{b}, a\right)$; their number can be two, one (the x-axis is a tangent line) or none, depending on whether $b^2 - 4ac > 0,\ = 0$ or < 0 holds.

For $b = 0$ the function (2.47) becomes the function $y = a + \frac{c}{x^2}$ (see **(Fig. 2.21)** the reciprocal power), and for $c = 0$ it becomes the homographic function $y = \frac{ax + b}{x}$, as a special case of (2.46).

2.4.4 Curves of Third Degree, Type II

The graph of the function

$$y = \frac{1}{ax^2 + bx + c} \quad (a \neq 0) \tag{2.48}$$

is a *curve of third degree* (type II) which is symmetric about the vertical line $x = -\frac{b}{2a}$ and the x-axis is its asymptote **(Fig. 2.19)**, because $\lim\limits_{x \to \pm\infty} y = 0$. Its shape depends on the signs of a and $\Delta = 4ac - b^2$. From the two cases $a > 0$ and $a < 0$ only the first one is considered, because reflecting the curve of $y = \frac{1}{(-a)x^2 - bx - c}$ with respect to the x-axis one gets the second one.

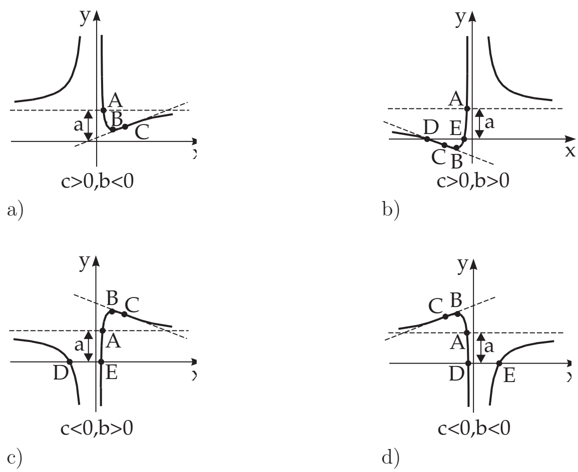


Figure 2.18

Case a) $\boldsymbol{\Delta > 0}$: The function is positive and continuous for arbitrary values of x and it is increasing on the interval $(-\infty\,, -\dfrac{b}{2a})$. Here it takes its maximum, $\dfrac{4a}{\Delta}$, then it is decreasing again in the interval $(-\dfrac{b}{2a}\,, \infty)$. The extreme point A of the curve is at $\left(-\dfrac{b}{2a}\,, \dfrac{4a}{\Delta}\right)$, the inflection points B and C are at $\left(-\dfrac{b}{2a} \pm \dfrac{\sqrt{\Delta}}{2a\sqrt{3}}\,, \dfrac{3a}{\Delta}\right)$; and for the corresponding slopes of the tangent lines (*angular coefficients*) we get $\tan\varphi = \mp a^2 \left(\dfrac{3}{\Delta}\right)^{3/2}$ **(Fig. 2.19a)**.

Case b) $\boldsymbol{\Delta = 0}$: The function is positive for arbitrary values of x, its value rises from 0 to $+\infty$, at $x = -\dfrac{b}{2a} = x_0$ it has a point of discontinuity (a pole), where $\lim\limits_{x \to x_0} y = +\infty$. Then its value falls from here back to 0 **(Fig. 2.19b)**.

Case c) $\boldsymbol{\Delta < 0}$: The value of y rises from 0 to $+\infty$, at the point of discontinuity it jumps to $-\infty$, and rises to the maximum, then falls back to $-\infty$; at the other point of discontinuity it jumps to $+\infty$, then it falls to 0. The extreme point A of the curve is at $\left(-\dfrac{b}{2a}\,, \dfrac{4a}{\Delta}\right)$. The points of discontinuity are at $x = \dfrac{-b \pm \sqrt{-\Delta}}{2a}$ **(Fig. 2.19c)**.

2.4.5 Curves of Third Degree, Type III

The graph of the function

$$y = \frac{x}{ax^2 + bx + c} \quad (a \neq 0, b \neq 0, c \neq 0) \tag{2.49}$$

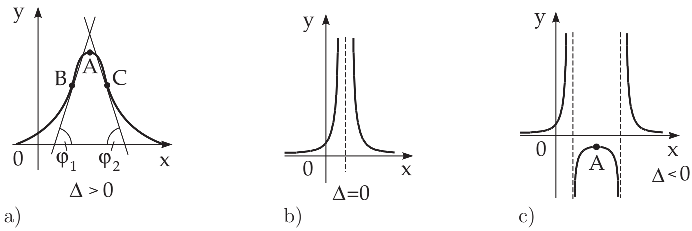


Figure 2.19

is a *curve of third degree* (type III) which goes through the origin, and has the x-axis **(Fig. 2.20)** as an asymptote. The behavior of the function depends on the signs of a and of $\Delta = 4ac - b^2$, and for $\Delta < 0$ also on the signs of the roots α and β of the equation $ax^2 + bx + c = 0$, and for $\Delta = 0$ also on the sign of b. From the two cases, $a > 0$ and $a < 0$, only the first one is considered because reflecting the curve of $y = \dfrac{x}{(-a)x^2 - bx - c}$ with respect to the x-axis yields the second one.

Case a) $\Delta > 0$: The function is continuous everywhere, its value falls from 0 to the minimum, then rises to the maximum, then falls again to 0.
The extreme points of the curve, A and B, are at $\left(\pm\sqrt{\dfrac{c}{a}},\ \dfrac{-b \pm 2\sqrt{ac}}{\Delta}\right)$; there are three inflection points **(Fig. 2.20a)**.

Case b) $\Delta = 0$: The behavior of the function depends on the sign of b, so there are two cases. In both cases there is a point of discontinuity at $x = -\dfrac{b}{2a}$; both curves have one inflection point.

• $b > 0$: The value of the function falls from 0 to $-\infty$, the function has a point of discontinuity, then the value of the function rises from $-\infty$ to the maximum, then decreases to 0 **(Fig. 2.20b$_1$)**. The extreme point A of the curve is at $A\left(+\sqrt{\dfrac{c}{a}}, \dfrac{1}{2\sqrt{ac} + b}\right)$.

• $b < 0$: The value of the function falls from 0 to the minimum, then rises to $+\infty$, running through the origin, then the function has a point of discontinuity, then the value of the function falls from $+\infty$ to 0 **(Fig. 2.20b$_2$)**. The extreme point A of the curve is at $A\left(-\sqrt{\dfrac{c}{a}}, -\dfrac{1}{2\sqrt{ac} - b}\right)$.

Case c) $\Delta < 0$: The function has two points of discontinuity, at $x = \alpha$ and $x = \beta$; its behavior depends on the signs of α and β.

• The signs of α and β are different: The value of the function falls from 0 to $-\infty$, jumps up to $+\infty$, then falls again from $+\infty$ to $-\infty$, running through the origin, then jumps again up to $+\infty$, then it falls tending to 0 **(Fig. 2.20c$_1$)**. The function has no extremum.

• The signs of α and β are both negative: The value of the function falls from 0 to $-\infty$, jumps up to $+\infty$, from here it goes through a minimum up to $+\infty$ again, jumps down to $-\infty$, then rises to a maximum, then falls tending to 0 **(Fig. 2.20c$_2$)**.
The extremum points A and B can be calculated with the same formula as in case a) of 2.4.5.

• The signs of α and β are both positive: The value of the function falls from 0 until the minimum, then rises to $+\infty$, jumps down to $-\infty$, then it rises to the maximum, then it falls again to $-\infty$, then jumps up to $+\infty$ and then it tends to 0 **(Fig. 2.20c$_3$)**.
The extremum points A and B can be calculated by the same formula as in case a) of 2.4.5.

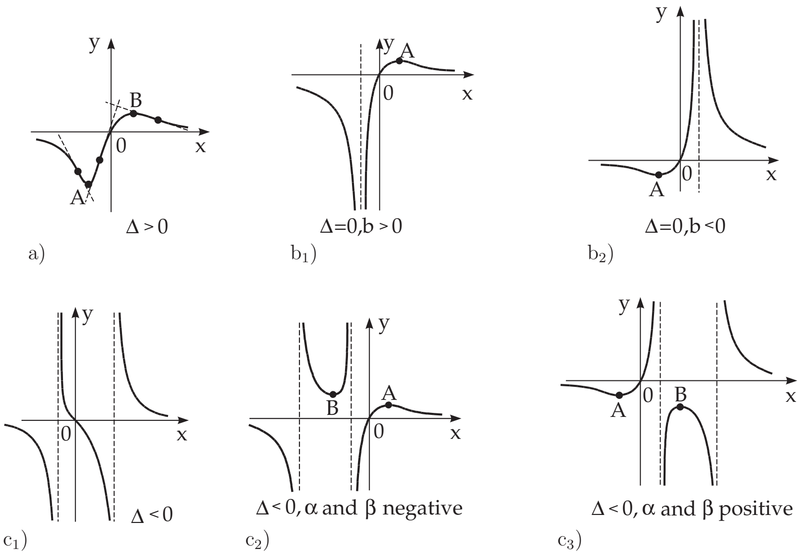


Figure 2.20

In all three cases the curve has one inflection point.

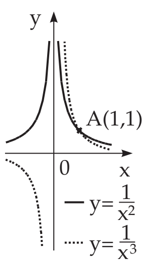


Figure 2.21

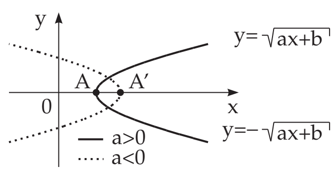


Figure 2.22

2.4.6 Reciprocal Powers

The graph of the function

$$y = \frac{a}{x^n} = ax^{-n} \quad (n > 0, \ \text{integer}; a \neq 0) \tag{2.50}$$

is a *curve of hyperbolic type* with the coordinate axes as asymptotes. The point of discontinuity is at $x = 0$ **(Fig. 2.21)**.

Case a) For $\boldsymbol{a > 0}$ and for even n the value of the function rises from 0 to $+\infty$, then it falls tending to 0, and it is always positive. For odd n it falls from 0 to $-\infty$, it jumps up to $+\infty$, then it falls tending to 0.

Case b) For $\boldsymbol{a < 0}$ and for even n the value of the function falls from 0 to $-\infty$, then it tends to 0, and it is always negative. For odd n it rises from 0 up to $+\infty$, jumps down to $-\infty$, then it tends to 0. The function does not have any extremum. The larger n is, the faster the curve approaches the x-axis, and the slower it approaches the y-axis. For even n the curve is symmetric with respect to the y-axis, for odd n it is center-symmetric and its center of symmetry is the origin. The **Fig. 2.21** shows the cases $n = 2$ and $n = 3$ for $a = 1$.

2.5 Irrational Functions

2.5.1 Square Root of a Linear Binomial

The union of the curve of the two functions

$$y = \pm\sqrt{ax+b} \quad (a \neq 0) \tag{2.51}$$

is a *parabola* with the x-axis as the symmetry axis. The vertex A is at $\left(-\frac{b}{a}, 0\right)$, the *semifocal chord* (see 3.5.2.10, p. 204) is $p = \frac{a}{2}$. The domain of the function and the shape of the curve depend on the sign of a **(Fig. 2.22)** (for more details about the parabola see 3.5.2.10, p. 204).

2.5.2 Square Root of a Quadratic Polynomial

The union of the graphs of the two functions

$$y = \pm\sqrt{ax^2+bx+c} \quad (a \neq 0,\ \Delta = 4ac - b^2 \neq 0) \tag{2.52}$$

is for $a < 0$ an *ellipse*, for $a > 0$ a *hyperbola* **(Fig. 2.23)**. One of the two symmetry axes is the x-axis, the other one is the line $x = -\frac{b}{2a}$.

The vertices A, C and B, D are at $\left(-\frac{b \pm \sqrt{-\Delta}}{2a}, 0\right)$ and $\left(-\frac{b}{2a}, \pm\sqrt{\frac{\Delta}{4a}}\right)$, where $\Delta = 4ac - b^2$.

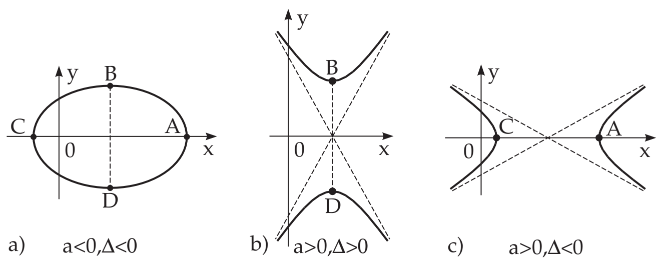


Figure 2.23

The domain of the function and the shape of the curve depend on the signs of a and Δ **(Fig. 2.23)**. For $a < 0$ and $\Delta > 0$ the function has only imaginary values, so no curve exists (for more details about the ellipse and hyperbola see 3.5.2.8, p. 199 and 3.5.2.9, p. 201).

2.5.3 Power Function

The power function

$$y = ax^k = ax^{\pm m/n} \quad (m,\ n \text{ integer, positive, coprime}) \tag{2.53}$$

is to be discussed for $k > 0$ and for $k < 0$ **(Fig. 2.24, Fig. 2.25)**. The investigation here can be restricted to the case $a = 1$, because for $a \neq 1$ the curve differs from the curve of $y = x^k$ only by a stretching in the direction of the y-axis by a factor $|a|$, and for a negative a by a reflection to the x-axis.

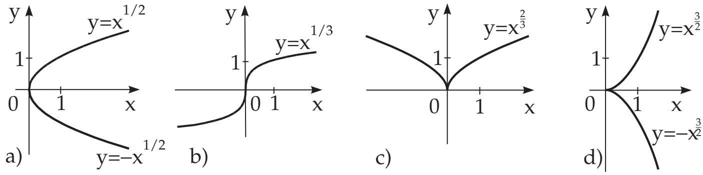


Figure 2.24

Case a) $\boldsymbol{k > 0,\ y = x^{m/n}}$: The shape of the curve is represented in four characteristic cases depending on the numbers m and n in **Fig. 2.24**. The curve goes through the points $(0,0)$ and $(1,1)$. For $k > 1$ the x-axis is a tangent line of the curve at the origin **(Fig. 2.24d)**, for $k < 1$ the y-axis is a tangent line also at the origin**(Fig. 2.24a,b,c)**. For even n the union of the graph of functions $y = \pm x^k$ may be considered: it has two branches symmetric to the x-axis **(Fig. 2.24a,d)**, for even m the curve is symmetric to the y-axis **(Fig. 2.24c)**. If m and n are both odd, the curve is symmetric with respect to the origin **(Fig. 2.24b)**. So the curves can have a vertex, a cusp or an inflection point at the origin **(Fig. 2.24)**. None of them has any asymptote.

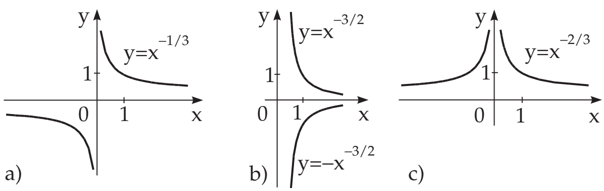


Figure 2.25

Case b) $\boldsymbol{k < 0,\ y = x^{-m/n}}$: The shape of the curve is represented in three characteristic cases depending on m and n in **Fig. 2.25**. The curve is a hyperbolic type curve, where the asymptotes coincide with the coordinate axes **(Fig. 2.25)**. The point of discontinuity is at $x = 0$. The greater $|k|$ is the faster the curve approaches the x-axis, and the slower it approaches the y-axis. The symmetry properties of the curves are the same as above for $k > 0$; they depend on whether m and n are even or odd. There is no extreme value.

2.6 Exponential Functions and Logarithmic Functions

2.6.1 Exponential Functions

The function

$$y = a^x = e^{bx} \quad (a > 0,\ \ b = \ln a)\,, \tag{2.54}$$

is called the *exponential function* and its graphical representation the *exponential curve* **(Fig. 2.26)**. From (2.54) for $a = e$ follows the function of the *natural exponential curve*

$$y = e^x. \tag{2.55}$$

The function has only positive values. Its domain is the interval $(-\infty, +\infty)$. For $a > 1$, i.e., for $b > 0$, the function is strictly monotone increasing and takes its values from 0 until ∞. For $a < 1$, i.e., for $b < 0$, it is strictly monotone decreasing, its value falls from ∞ until 0. The larger $|b|$ is, the greater is

the speed of growth and decay. The curve goes through the point $(0, 1)$ and approaches asymptotically the x-axis, for $b > 0$ on the right and for $b < 0$ on the left, and faster for greater values of $|b|$. The function $y = a^{-x} = \left(\frac{1}{a}\right)^x$ increases for $a < 1$ and decreases for $a > 1$.

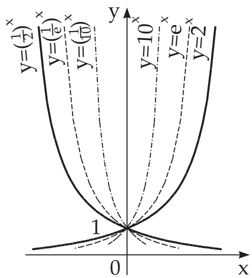


Figure 2.26

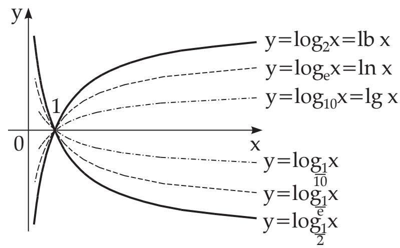


Figure 2.27

2.6.2 Logarithmic Functions

The function

$$y = \log_a x \quad (a > 0, \ \ a \neq 1) \tag{2.56}$$

gives the *logarithmic curve* **(Fig. 2.27)**; the curve is the reflection of the exponential curve with respect to the line $y = x$. From (2.56) for $a = e$ follows the curve of the *natural logarithm*

$$y = \ln x\,. \tag{2.57}$$

The real logarithmic function is defined only for $x > 0$. For $a > 1$ it is strictly monotone increasing and takes its values from $-\infty$ to $+\infty$, for $a < 1$ it is strictly monotone decreasing, and takes its values from $+\infty$ to $-\infty$, and the greater $|\ln a|$ is, the faster the growth and decay. The curve goes through the point $(1, 0)$ and approaches asymptotically the y-axis, for $a > 1$ down, for $a < 1$ up, and again faster for larger values of $|\ln a|$.

2.6.3 Error Curve

The function

$$y = e^{-(ax)^2} \tag{2.58}$$

gives the *error curve (Gauss error distribution curve)* **(Fig. 2.28)**. Since the function is even, the y-axis is the symmetry axis of the curve and the larger $|a|$ is, the faster it approaches asymptotically the x-axis. It takes its maximum at zero, and it is equal to one, so the extreme point A of the curve is at $(0, 1)$, the inflection points of the curve B, C are at $\left(\pm\frac{1}{a\sqrt{2}}, \frac{1}{\sqrt{e}}\right)$.

The slopes of the tangent lines are here $\tan\varphi = \mp a\sqrt{2/e}\,$.

A very important application of the *error curve* (2.58) is the description of the *normal distribution properties of the observational error* (see 16.2.4.1, p. 818.):

$$y = \varphi(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{x^2}{2\sigma^2}\right). \tag{2.59}$$

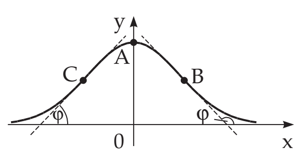


Figure 2.28

2.6.4 Exponential Sum

The function

$$y = ae^{bx} + ce^{dx} \tag{2.60}$$

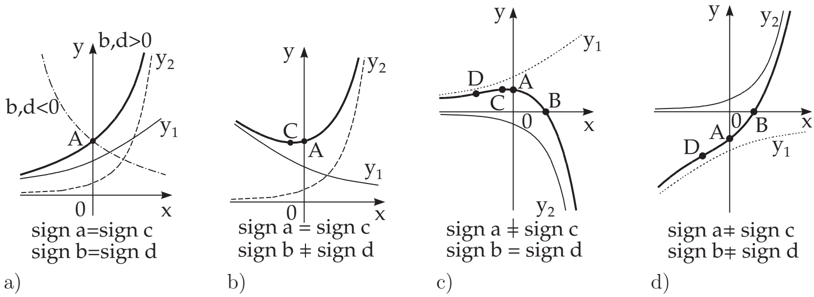


Figure 2.29

is represented in **Fig. 2.29** for the characteristic sign relations. The sum of the functions is got by adding the ordinates of the curves, i.e., the summands are $y_1 = ae^{bx}$ and $y_2 = ce^{dx}$. The function is continuous. If none of the numbers a, b, c, d is equal to 0, the curve has one of the four forms represented in **Fig. 2.29**. Depending on the signs of the parameters it is possible, that the graphs are reflected over a coordinate axis.

The intersection points A and B of the curve with the y-axis and with the x-axis are at $(0, a + c)$, and at $\left(\frac{\ln(-a/c)}{d-b}, 0\right)$ respectively, the extremum C is at $x = \frac{1}{d-b} \ln\left(-\frac{ab}{cd}\right)$, and the inflection point D is at $x = \frac{1}{d-b} \ln\left(-\frac{ab^2}{cd^2}\right)$, in the case when they exist.

Case a) The parameters a and c, and b and d have the same signs: The function does not change its sign, it is strictly monotone; its value is changing from 0 to $+\infty$ or to $-\infty$ or it is changing from $+\infty$ or from $-\infty$ to 0. There is no inflection point. The asymptote is the x-axis **(Fig. 2.29a)**.

Case b) The parameters a and c have the same sign, b and d have different signs: The function does not change its sign and either comes from $+\infty$ and arrives at $+\infty$ and has a minimum or comes from $-\infty$, goes to $-\infty$ and has a maximum. There is no inflection point **(Fig. 2.29b)**.

Case c) The parameters a and c have different signs, b and d have the same signs: The function has one extremum and it is strictly monotone before and after. It changes its sign once. Its value changes whether from 0 until the extremum, then goes to $+\infty$ or $-\infty$ or it comes first from $+\infty$ or $-\infty$, takes the extremum, then approaches 0. The x-axis is an asymptote, the extreme point of the curve is at C and the inflection point at D **(Fig. 2.29c)**.

Case d) The parameters a and c and also b and d have different signs: The function is strictly monotone, its value rises from $-\infty$ to $+\infty$ or it falls from $+\infty$ to $-\infty$. It has an inflection point D **(Fig. 2.29d)**.

2.6.5 Generalized Error Function

The curve of the function

$$y = a\exp(bx + cx^2) = a\exp\left(-\frac{b^2}{4c^2}\right)\exp\left(c\left(x + \frac{b}{2c}\right)^2\right) \quad (c \neq 0) \tag{2.61}$$

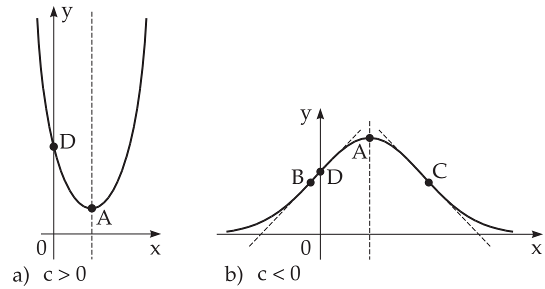


Figure 2.30

can be considered as the generalization of the error function (2.58); it results in a symmetric curve with respect to the vertical line $x = -\frac{b}{2c}$, it has no intersection point with the x-axis, and the intersection point D with the y-axis is at $(0, a)$ **(Fig. 2.30a,b)**.

The shape of the curve depends on the signs of a and c. Here only the case $a > 0$ is discussed, because the curve for $a < 0$ is got by reflecting it in the x-axis.

Case a) $c > 0$: The value of the function falls from $+\infty$ until the minimum, and then rises again to $+\infty$. It is always positive. The extreme point A of the curve is at $\left(-\frac{b}{2c}, a\exp\left(\frac{b^2}{4c}\right)\right)$ and it corresponds to the minimum of the function; there is no inflection point or asymptote **(Fig. 2.30a)**.

Case b) $c < 0$: The x-axis is the asymptote. The extreme point A of the curve is at $\left(-\frac{b}{2c}, a\exp\left(-\frac{b^2}{4c}\right)\right)$ and it corresponds to the maximum of the function. The inflection points B and C are at
$\left(\frac{-b \pm \sqrt{-2c}}{2c}, a\exp\left(\frac{-(b^2+2c)}{4c}\right)\right)$ **(Fig 2.30b)**.

2.6.6 Product of Power and Exponential Functions

The function

$$y = ax^b e^{cx} \tag{2.62}$$

is discussed here only in the case $a > 0$, because in the case $a < 0$ the curve is got by reflecting it in the x-axis. For a non-integer b the function is defined only for $x > 0$, and for an integer b the shape of the curve for negative x can be deduced also from the following cases **(Fig. 2.31)**.

Fig. 2.31 shows how the curve behaves for arbitrary parameters.

For $b > 0$ the curve passes through the origin. The tangent line at this point for $b > 1$ is the x-axis, for $b = 1$ the line $y = x$, for $0 < b < 1$ the y-axis. For $b < 0$ the y-axis is an asymptote. For $c > 0$ the function is increasing and exceeds any value, for $c < 0$ it tends asymptotically to 0. For different signs of b and c the function has an extremum at $x = -\frac{b}{c}$ (point A on the curve). The curve has either no or one or two inflection points at $x = -\frac{b \pm \sqrt{b}}{c}$ (points C and D see **Fig. 2.31c,e,f,g**).

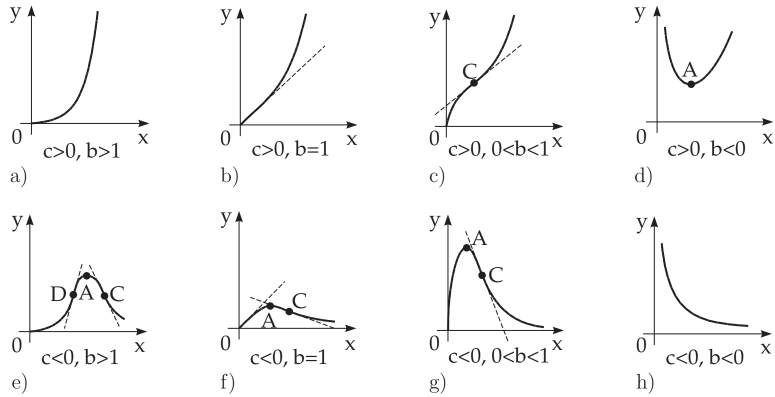


Figure 2.31

2.7 Trigonometric Functions (Functions of Angles)

2.7.1 Basic Notions

2.7.1.1 Definition and Representation

1. Definition

The trigonometric functions are introduced by geometric considerations. So in their definition and also in their arguments degree or radian measure is used (see 3.1.1.5, p. 131).

2. Sine

The *standard sine function*

$$y = \sin x \tag{2.63}$$

is a continuous curve with period $T = 2\pi$ (see **Fig. 2.32a**).

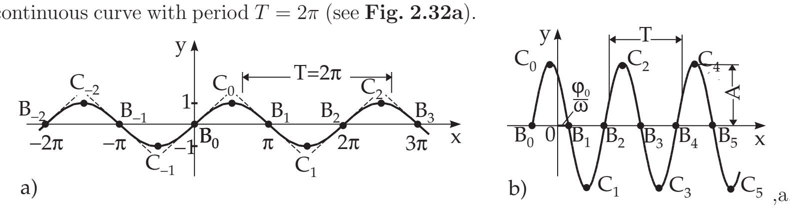


Figure 2.32

The intersection points B_0, B_1, B_{-1}, B_2, B_{-2},. . . , with $B_k = (k\pi, 0)$ $(k = 0, \pm1, \pm2, \ldots)$ of the standard sine curve and the x-axis are the inflection points of the curve. Here the angle of slope of the tangent line with the x axis is $\pm\dfrac{\pi}{4}$. The extreme points of the curve are at C_0, C_1, C_{-1}, C_2, C_{-2}, . . . with $C_k = \left((k + \frac{1}{2})\pi, (-1)^k\right)$ $(k = 0, \pm1, \pm2, \ldots\Big)$. For every value of the function y there is $-1 \le y \le 1$.

The *general sine function*

$$y = A\sin(\omega x + \varphi_0) \tag{2.64}$$

with an amplitude $|A|$, frequency ω, and phase shift φ_0 is represented in **Fig. 2.32b**.
Comparing the standard and the general sine curve **(Fig. 2.32b)** it can be seen that in the general case the curve is stretched in the direction of y by a factor $|A|$, in the direction of x it is compressed by a factor $\frac{1}{\omega}$, and it is shifted to the left by a segment $\frac{\varphi_0}{\omega}$. The period is $T = \frac{2\pi}{\omega}$. The intersection points with the x-axis are $B_k = \left(\frac{k\pi - \varphi_0}{\omega}, 0\right)$ $(k = 0, \pm 1, \pm 2, \ldots)$. The extreme points are $C_k = \left(\frac{\left[\left(k + \frac{1}{2}\right)\pi - \varphi_0\right]}{\omega}, (-1)^k A\right)$ $(k = 0, \pm 1, \pm 2, \ldots)$.

3. Cosine

The *standard cosine function*

$$y = \cos x = \sin(x + \frac{\pi}{2}) \tag{2.65}$$

is represented in **Fig. 2.33**.

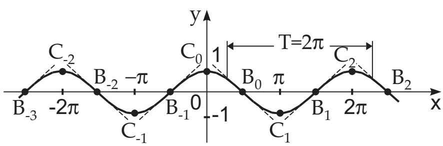


Figure 2.33

The intersection points with the x-axis
$B_0, B_1, B_2, \ldots, B_k = \left(\left(k + \frac{1}{2}\right)\pi, 0\right)$ $(k = 0, \pm 1, \pm 2, \ldots)$ are also the inflection points. The angle of slope of the tangent line is $\pm\frac{\pi}{4}$.
The extreme points are $C_0, C_1, \ldots, C_k = (k\pi, (-1)^k)$ $(k = 0, \pm 1, \pm 2, \ldots)$.
The *general cosine function*

$$y = A\cos(\omega x + \varphi_0) \tag{2.66a}$$

can be transformed into the form

$$y = A\sin\left(\omega x + \varphi_0 + \frac{\pi}{2}\right), \tag{2.66b}$$

i.e., the general sine function shifted left by $\varphi = \frac{\pi}{2}$.

4. Tangent

The *tangent function*

$$y = \tan x = \frac{\sin x}{\cos x} \tag{2.67}$$

has period $T = \pi$ and the asymptotes are $x = \left(k + \frac{1}{2}\right)\pi$ $(k = 0, \pm 1, \pm 2, \ldots)$ **(Fig. 2.34)**. The function is monotone increasing in the intervals $\left(-\frac{\pi}{2} + k\pi, +\frac{\pi}{2} + k\pi\right)$ $(k = 0, \pm 1, \pm 2, \ldots)$ and takes values from $-\infty$ to $+\infty$. The curve has intersection points with the x-axis at A_0, A_1, A_{-1}, A_2, $A_{-2}, \ldots$, $A_k = (k\pi, 0)$ $(k = 0, \pm 1, \pm 2, \ldots)$, these points are the inflection points and the angle of slope of the tangent line is $\frac{\pi}{4}$.

5. Cotangent

The *cotangent function*

$$y = \cot x = \frac{\cos x}{\sin x} = \frac{1}{\tan x} = -\tan\left(x + \frac{\pi}{2}\right) \tag{2.68}$$

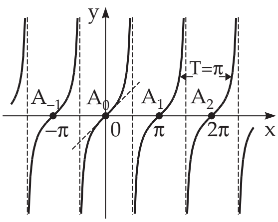


Figure 2.34

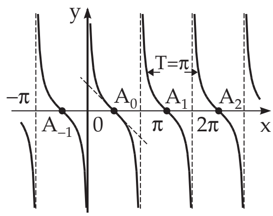


Figure 2.35

has a graph which is the tangent curve reflected with respect to the x-axis and shifted to the left by $\frac{\pi}{2}$ **(Fig. 2.35)**. The asymptotes are $x = k\pi$ $(k = 0, \pm 1, \pm 2, \ldots)$. Between 0 and π the function is monotone decreasing and takes its values from $+\infty$ until $-\infty$; the function has period $T = \pi$. The intersection points with the x-axis are at A_0, A_1, A_{-1}, A_2, $A_{-2}, \ldots$ with $A_k = \left(\left(k + \frac{1}{2}\right)\pi, 0\right)$ $(k = 0, \pm 1, \pm 2, \ldots)$, they are the inflection points of the curve and here the angle of the tangent line is $-\frac{\pi}{4}$.

6. Secant

The *secant function*

$$y = \sec x = \frac{1}{\cos x} \tag{2.69}$$

has period $T = 2\pi$, the asymptotes are $x = \left(k + \frac{1}{2}\right)\pi$ $(k = 0, \pm 1, \pm 2, \ldots)$; and obviously $|y| \geq 1$ holds. The extreme points corresponding to the maxima of the function are A_0, A_1, $A_{-1}, \ldots$

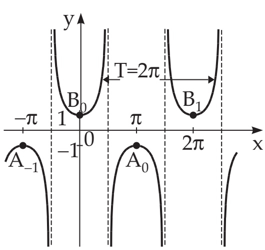


Figure 2.36

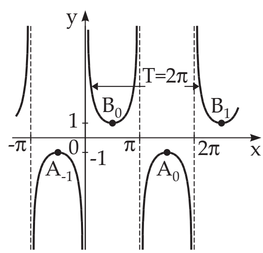


Figure 2.37

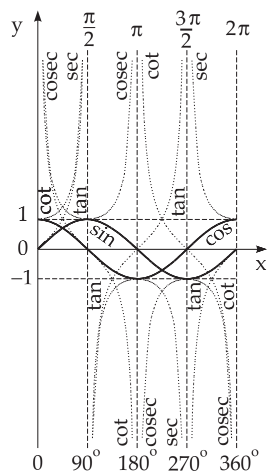


Figure 2.38

with $A_k = ((2k+1)\pi, -1)$ $(k = 0, \pm 1, \pm 2, \ldots)$, the extreme points corresponding to the minima of the function are B_0, B_1, $B_{-1}, \ldots$ with $B_k = (2k\pi, +1)$ $(k = 0, \pm 1, \pm 2, \ldots)$ **(Fig. 2.36)**.

7. Cosecant

The *cosecant function*

$$y = \operatorname{cosec} x = \frac{1}{\sin x} \tag{2.70}$$

has a graph which is the graph of the secant shifted to the right by $x = \frac{\pi}{2}$. The asymptotes are $x = k\pi$ $(k = 0, \pm 1, \pm 2, \ldots)$. The extreme points corresponding to the maxima of the function are A_0, A_1, $A_{-1}, \ldots$ with $A_k = \left(\frac{4k+3}{2}\pi, -1\right)$ $(k = 0, \pm 1, \pm 2, \ldots)$ and the the points corresponding to the minima of the function are B_0, B_1, $B_{-1}, \ldots$ with $B_k = \left(\frac{4k+1}{2}\pi, +1\right)$ $(k = 0, \pm 1, \pm 2, \ldots)$ **(Fig. 2.37)**.

2.7.1.2 Range and Behavior of the Functions

1. Angle Domain $0 \leq x \leq 360°$

The six trigonometric functions are represented together in **Fig. 2.38** in all the four quadrants for a complete domain of angles from 0° to 360° or for a complete domain of radians from 0 to 2π.
In **Table 2.1** there is a review of the domain and the range of these functions. The signs of the functions depend on the quadrant where the argument is taken from, and these are reviewed in **Table 2.2**.

Table 2.1 Domain and range of trigonometric functions

Domain	Range	Domain	Range
$-\infty < x < \infty$	$\left\{\begin{array}{l} -1 \leq \sin x \leq 1 \\ -1 \leq \cos x \leq 1 \end{array}\right.$	$x \neq (2k+1)\frac{\pi}{2}$ $x \neq k\pi$ $(k = 0, \pm 1, \pm 2, \ldots)$	$-\infty < \tan x < \infty$ $-\infty < \cot x < \infty$

2. Function Values for Some Special Arguments (see Table 2.3)

3. Arbitrary Angle

Since the trigonometric functions are periodic (period 360° or 180°), the determination of their values for an arbitrary argument x can be reduced by the following rules.
Argument $x > 360°$ or $x > 180°$: If the angle is greater than 360° or greater than 180°, then it is to be reduced for a value α, for which $0 \leq \alpha \leq 360°$ or $0 \leq \alpha \leq 180°$ holds, in the following way (n integer):

$$\sin(360° \cdot n + \alpha) = \sin\alpha\,, \qquad (2.71) \qquad\qquad \cos(360° \cdot n + \alpha) = \cos\alpha\,, \qquad (2.72)$$

$$\tan(180° \cdot n + \alpha) = \tan\alpha\,, \qquad (2.73) \qquad\qquad \cot(180° \cdot n + \alpha) = \cot\alpha\,. \qquad (2.74)$$

Table 2.2 Signs of trigonometric functions

Quadrant	Angle	sin	cos	tan	cot	sec	csc
I	from 0° to 90°	+	+	+	+	+	+
II	from 90° to 180°	+	−	−	−	−	+
III	from 180° to 270°	−	−	+	+	−	−
IV	from 270° to 360°	−	+	−	−	+	−

Argument $x < 0$: If the argument is negative, then the following formulas reduce the calculations to functions for positive argument:

$$\sin(-\alpha) = -\sin\alpha\,, \qquad (2.75) \qquad\qquad \cos(-\alpha) = \cos\alpha\,, \qquad (2.76)$$

$$\tan(-\alpha) = -\tan\alpha\,, \qquad (2.77) \qquad\qquad \cot(-\alpha) = -\cot\alpha\,. \qquad (2.78)$$

Table 2.3 Values of trigonometric functions for $0°, 30°, 45°, 60°$ and $90°$.

Angle	Radian	sin	cos	tan	cot	sec	csc
$0°$	0	0	1	0	$\mp\infty$	1	$\mp\infty$
$30°$	$\frac{1}{6}\pi$	$\frac{1}{2}$	$\frac{\sqrt{3}}{2}$	$\frac{\sqrt{3}}{3}$	$\sqrt{3}$	$\frac{2\sqrt{3}}{3}$	2
$45°$	$\frac{1}{4}\pi$	$\frac{\sqrt{2}}{2}$	$\frac{\sqrt{2}}{2}$	1	1	$\sqrt{2}$	$\sqrt{2}$
$60°$	$\frac{1}{3}\pi$	$\frac{\sqrt{3}}{2}$	$\frac{1}{2}$	$\sqrt{3}$	$\frac{\sqrt{3}}{3}$	2	$\frac{2\sqrt{3}}{3}$
$90°$	$\frac{1}{2}\pi$	1	0	$\pm\infty$	0	$\pm\infty$	1

Table 2.4 Reduction formulas and quadrant relations of trigonometric functions

Function	$x = 90° \pm \alpha$	$x = 180° \pm \alpha$	$x = 270° \pm \alpha$	$x = 360° - \alpha$
$\sin x$	$+\cos\alpha$	$\mp\sin\alpha$	$-\cos\alpha$	$-\sin\alpha$
$\cos x$	$\mp\sin\alpha$	$-\cos\alpha$	$\pm\sin\alpha$	$+\cos\alpha$
$\tan x$	$\mp\cot\alpha$	$\pm\tan\alpha$	$\mp\cot\alpha$	$-\tan\alpha$
$\cot x$	$\mp\tan\alpha$	$\pm\cot\alpha$	$\mp\tan\alpha$	$-\cot\alpha$

Argument x for $90° < x < 360°$: If $90° < x < 360°$ holds, then the arguments are to be reduced for an acute angle α by the *reduction formulas* given in **Table 2.4**. The relations between the values of the functions belonging to the arguments which differ from each other by $90°, 180°$ or $270°$ or which complete each other to $90°, 180°$ or $270°$ are called *quadrant relations.*
The first and second columns of **Table 2.4** give the *complementary angle formulas*, and the first and third ones give the *supplementary angle formulas*. Because $x = 90° - \alpha$ is the complementary angle (see 3.1.1.2, p. 130) of α, the following relations

$$\cos\alpha = \sin x = \sin(90° - \alpha)\,, \quad (2.79a) \qquad \sin\alpha = \cos x = \cos(90° - \alpha) \quad (2.79b)$$

are called the *complementary angle formulas.*

For $\alpha + x = 180°$ the relations between the trigonometric functions for supplementary angles (see 3.1.1.2, p. 130)

$$\sin\alpha = \sin x = \sin(180° - \alpha)\,, \quad (2.80a) \qquad -\cos\alpha = \cos x = \cos(180° - \alpha) \quad (2.80b)$$

are called *supplementary angle formulas.*
Argument x for $0° < x < 90°$: The values of trigonometric functions for acute angles ($0° < x < 90°$) have been taken formerly from tables, today calculators are used.

■ $\sin(-1000°) = -\sin 1000° = -\sin(360° \cdot 2 + 280°) = -\sin 280° = +\cos 10° = +0.9848.$

4. Angles in Radian Measure

The arguments given in radian measure, i.e., in units of radians, can be easily converted by formula (3.2) (see 3.1.1.5, p. 131).

2.7.2 Important Formulas for Trigonometric Functions

Remark: Trigonometric functions with complex argument z are discussed in 14.5.2, p. 759.

2.7.2.1 Relations Between the Trigonometric Functions

$$\sin^2\alpha + \cos^2\alpha = 1\,, \tag{2.81}$$

$$\sec^2\alpha - \tan^2\alpha = 1, \tag{2.82}$$

$$\mathrm{cosec}^2\,\alpha - \cot^2\alpha = 1\,, \tag{2.83}$$

$$\sin\alpha \cdot \mathrm{cosec}\,\alpha = 1\,, \tag{2.84}$$

$$\cos\alpha \cdot \sec\alpha = 1\,, \tag{2.85}$$

$$\tan\alpha \cdot \cot\alpha = 1\,, \tag{2.86}$$

$$\frac{\sin\alpha}{\cos\alpha} = \tan\alpha\,, \tag{2.87}$$

$$\frac{\cos\alpha}{\sin\alpha} = \cot\alpha\,. \tag{2.88}$$

Some important relations are summarized in **Table 2.5** for $0 < \alpha < \pi/2$ in order to create an easy survey. For other intervals in **Table 2.5** the square roots are always considered with the sign which corresponds to the quadrant where the argument is.

2.7.2.2 Trigonometric Functions of the Sum and Difference of Two Angles (Addition Theorems)

$$\sin(\alpha\pm\beta) = \sin\alpha\cos\beta\pm\cos\alpha\sin\beta, \tag{2.89}$$

$$\cos(\alpha\pm\beta) = \cos\alpha\cos\beta\mp\sin\alpha\sin\beta, \tag{2.90}$$

$$\tan(\alpha\pm\beta) = \frac{\tan\alpha\pm\tan\beta}{1\mp\tan\alpha\tan\beta}, \tag{2.91}$$

$$\cot(\alpha\pm\beta) = \frac{\cot\alpha\cot\beta\mp 1}{\cot\beta\pm\cot\alpha}, \tag{2.92}$$

$$\begin{aligned}\sin(\alpha+\beta+\gamma) &= \sin\alpha\cos\beta\cos\gamma + \cos\alpha\sin\beta\cos\gamma \\ &\quad + \cos\alpha\cos\beta\sin\gamma - \sin\alpha\sin\beta\sin\gamma,\end{aligned} \tag{2.93}$$

$$\begin{aligned}\cos(\alpha+\beta+\gamma) &= \cos\alpha\cos\beta\cos\gamma - \sin\alpha\sin\beta\cos\gamma \\ &\quad - \sin\alpha\cos\beta\sin\gamma - \cos\alpha\sin\beta\sin\gamma.\end{aligned} \tag{2.94}$$

2.7.2.3 Trigonometric Functions of an Integer Multiple of an Angle

$$\sin 2\alpha = 2\sin\alpha\cos\alpha, \tag{2.95}$$

$$\sin 3\alpha = 3\sin\alpha - 4\sin^3\alpha, \tag{2.96}$$

$$\cos 2\alpha = \cos^2\alpha - \sin^2\alpha, \tag{2.97}$$

$$\cos 3\alpha = 4\cos^3\alpha - 3\cos\alpha, \tag{2.98}$$

$$\sin 4\alpha = 8\cos^3\alpha\sin\alpha - 4\cos\alpha\sin\alpha, \tag{2.99}$$

$$\cos 4\alpha = 8\cos^4\alpha - 8\cos^2\alpha + 1, \tag{2.100}$$

$$\tan 2\alpha = \frac{2\tan\alpha}{1-\tan^2\alpha}, \tag{2.101}$$

$$\tan 3\alpha = \frac{3\tan\alpha - \tan^3\alpha}{1-3\tan^2\alpha}, \tag{2.102}$$

$$\tan 4\alpha = \frac{4\tan\alpha - 4\tan^3\alpha}{1-6\tan^2\alpha+\tan^4\alpha,} \tag{2.103}$$

$$\cot 2\alpha = \frac{\cot^2\alpha - 1}{2\cot\alpha}, \tag{2.104}$$

$$\cot 3\alpha = \frac{\cot^3\alpha - 3\cot\alpha}{3\cot^2\alpha - 1}, \tag{2.105}$$

$$\cot 4\alpha = \frac{\cot^4\alpha - 6\cot^2\alpha + 1}{4\cot^3\alpha - 4\cot\alpha}. \tag{2.106}$$

For larger values of n in order to gain a formula for $\sin n\alpha$ and $\cos n\alpha$ the de Moivre formula is to be used (see 1.5.3.5, p. 38).

Using the binomial theorem (see 1.1.6.4, p. 12) gives:

$$\cos n\alpha + \mathrm{i}\sin n\alpha = \sum_{k=0}^{n}\binom{n}{k}\mathrm{i}^k\cos^{n-k}\alpha\sin^k\alpha = (\cos\alpha + \mathrm{i}\sin\alpha)^n$$
$$= \cos^n\alpha + \mathrm{i}n\cos^{n-1}\alpha\sin\alpha$$
$$-\binom{n}{2}\cos^{n-2}\alpha\sin^2\alpha - \mathrm{i}\binom{n}{3}\cos^{n-3}\alpha\sin^3\alpha + \binom{n}{4}\cos^{n-4}\alpha\sin^4\alpha + \ldots . \quad (2.107)$$

With this it follows:

$$\cos n\alpha = \cos^n\alpha - \binom{n}{2}\cos^{n-2}\alpha\sin^2\alpha + \binom{n}{4}\cos^{n-4}\alpha\sin^4\alpha - \binom{n}{6}\cos^{n-6}\alpha\sin^6\alpha + \ldots , \quad (2.108)$$

$$\sin n\alpha = n\cos^{n-1}\alpha\sin\alpha - \binom{n}{3}\cos^{n-3}\alpha\sin^3\alpha + \binom{n}{5}\cos^{n-5}\alpha\sin^5\alpha - \ldots . \quad (2.109)$$

Table 2.5 Relations between the trigonometric functions of the same argument in the interval $0 < \alpha < \frac{\pi}{2}$

α	$\sin\alpha$	$\cos\alpha$	$\tan\alpha$	$\cot\alpha$
$\sin\alpha$	–	$\sqrt{1-\cos^2\alpha}$	$\frac{\tan\alpha}{\sqrt{1+\tan^2\alpha}}$	$\frac{1}{\sqrt{1+\cot^2\alpha}}$
$\cos\alpha$	$\sqrt{1-\sin^2\alpha}$	–	$\frac{1}{\sqrt{1+\tan^2\alpha}}$	$\frac{\cot\alpha}{\sqrt{1+\cot^2\alpha}}$
$\tan\alpha$	$\frac{\sin\alpha}{\sqrt{1-\sin^2\alpha}}$	$\frac{\sqrt{1-\cos^2\alpha}}{\cos\alpha}$	–	$\frac{1}{\cot\alpha}$
$\cot\alpha$	$\frac{\sqrt{1-\sin^2\alpha}}{\sin\alpha}$	$\frac{\cos\alpha}{\sqrt{1-\cos^2\alpha}}$	$\frac{1}{\tan\alpha}$	–

2.7.2.4 Trigonometric Functions of Half-Angles

In the following formulas the sign of the square root must be chosen positive or negative, according to the quadrant where the half-angle is.

$$\sin\frac{\alpha}{2} = \sqrt{\frac{1}{2}(1-\cos\alpha)}, \quad (2.110) \qquad \cos\frac{\alpha}{2} = \sqrt{\frac{1}{2}(1+\cos\alpha)}, \quad (2.111)$$

$$\tan\frac{\alpha}{2} = \sqrt{\frac{1-\cos\alpha}{1+\cos\alpha}} = \frac{1-\cos\alpha}{\sin\alpha} = \frac{\sin\alpha}{1+\cos\alpha}, \quad (2.112)$$

$$\cot\frac{\alpha}{2} = \sqrt{\frac{1+\cos\alpha}{1-\cos\alpha}} = \frac{1+\cos\alpha}{\sin\alpha} = \frac{\sin\alpha}{1-\cos\alpha}. \quad (2.113)$$

2.7.2.5 Sum and Difference of Two Trigonometric Functions

$$\sin\alpha+\sin\beta=2\sin\frac{\alpha+\beta}{2}\cos\frac{\alpha-\beta}{2}, \quad (2.114)$$

$$\sin\alpha-\sin\beta=2\cos\frac{\alpha+\beta}{2}\sin\frac{\alpha-\beta}{2}, \quad (2.115)$$

$$\cos\alpha+\cos\beta=2\cos\frac{\alpha+\beta}{2}\cos\frac{\alpha-\beta}{2}, \quad (2.116)$$

$$\cos\alpha-\cos\beta=-2\sin\frac{\alpha+\beta}{2}\sin\frac{\alpha-\beta}{2}, \quad (2.117)$$

$$\tan\alpha\pm\tan\beta=\frac{\sin(\alpha\pm\beta)}{\cos\alpha\cos\beta}, \quad (2.118)$$

$$\cot\alpha\pm\cot\beta=\pm\frac{\sin(\alpha\pm\beta)}{\sin\alpha\sin\beta}, \quad (2.119)$$

$$\tan\alpha+\cot\beta=\frac{\cos(\alpha-\beta)}{\cos\alpha\sin\beta}, \quad (2.120)$$

$$\cot\alpha-\tan\beta=\frac{\cos(\alpha+\beta)}{\sin\alpha\cos\beta}. \quad (2.121)$$

2.7.2.6 Products of Trigonometric Functions

$$\sin\alpha\sin\beta=\frac{1}{2}[\cos(\alpha-\beta)-\cos(\alpha+\beta)], \quad (2.122)$$

$$\cos\alpha\cos\beta=\frac{1}{2}[\cos(\alpha-\beta)+\cos(\alpha+\beta)], \quad (2.123)$$

$$\sin\alpha\cos\beta=\frac{1}{2}[\sin(\alpha-\beta)+\sin(\alpha+\beta)], \quad (2.124)$$

$$\sin\alpha\sin\beta\sin\gamma=\frac{1}{4}[\sin(\alpha+\beta-\gamma)+\sin(\beta+\gamma-\alpha)+\sin(\gamma+\alpha-\beta)-\sin(\alpha+\beta+\gamma)], \quad (2.125)$$

$$\sin\alpha\cos\beta\cos\gamma=\frac{1}{4}[\sin(\alpha+\beta-\gamma)-\sin(\beta+\gamma-\alpha)+\sin(\gamma+\alpha-\beta)+\sin(\alpha+\beta+\gamma)], \quad (2.126)$$

$$\sin\alpha\sin\beta\cos\gamma=\frac{1}{4}[-\cos(\alpha+\beta-\gamma)+\cos(\beta+\gamma-\alpha)+\cos(\gamma+\alpha-\beta)-\cos(\alpha+\beta+\gamma)], \quad (2.127)$$

$$\cos\alpha\cos\beta\cos\gamma=\frac{1}{4}[\cos(\alpha+\beta-\gamma)+\cos(\beta+\gamma-\alpha)+\cos(\gamma+\alpha-\beta)+\cos(\alpha+\beta+\gamma)]. \quad (2.128)$$

2.7.2.7 Powers of Trigonometric Functions

$$\sin^2\alpha=\frac{1}{2}(1-\cos 2\alpha), \quad (2.129)$$

$$\cos^2\alpha=\frac{1}{2}(1+\cos 2\alpha), \quad (2.130)$$

$$\sin^3\alpha=\frac{1}{4}(3\sin\alpha-\sin 3\alpha), \quad (2.131)$$

$$\cos^3\alpha=\frac{1}{4}(\cos 3\alpha+3\cos\alpha), \quad (2.132)$$

$$\sin^4\alpha=\frac{1}{8}(\cos 4\alpha-4\cos 2\alpha+3), \quad (2.133)$$

$$\cos^4\alpha=\frac{1}{8}(\cos 4\alpha+4\cos 2\alpha+3). \quad (2.134)$$

For large values of n $\sin^n \alpha$ and $\cos^n \alpha$, can be expressd by applying the formulas for $\cos n\alpha$ and $\sin n\alpha$ (see 2.7.2.3, p. 82).

2.7.3 Description of Oscillations

2.7.3.1 Formulation of the Problem

In engineering and physics one often meets quantities depending on time and given in the form

$$u(t) = A \sin(\omega t + \varphi). \tag{2.135}$$

They are called also *sinusoidal quantities*. Their dependence on time results in a *harmonic oscillation*. The graphical representation of (2.135) results in a *general sine curve*, as shown in **Fig. 2.39**.

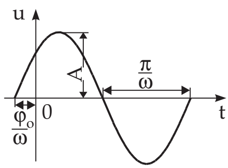


Figure 2.39

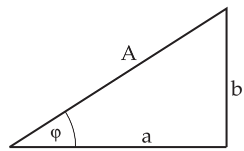


Figure 2.40

The general sine curve differs from the simple sine curve $y = \sin x$:

a) by the *amplitude* A, i.e., the greatest distance between its points and the time axis t,

b) by the *period* $T = \dfrac{2\pi}{\omega}$, which corresponds to the *wavelength* (with ω as the *frequency of the oscillation*, which is called the *angular* or *radial frequency* in wave theory),

c) by the *initial phase* or *phase shift* by the initial angle $\varphi \neq 0$.

The quantity $u(t)$ can also be written in the form

$$u(t) = a \sin \omega t + b \cos \omega t. \tag{2.136}$$

Here for a and b $A = \sqrt{a^2 + b^2}$ and $\tan \varphi = \dfrac{b}{a}$ holds. The quantities a, b, A and φ can be represented as sides and angle of a right triangle **(Fig. 2.40)**.

2.7.3.2 Superposition of Oscillations

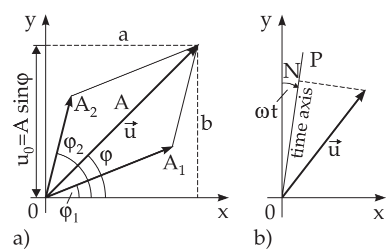


Figure 2.41

In the simplest case the *superposition of oscillations* is the *addition of two oscillations* with the same frequency. It results again in a harmonic oscillation with the same frequency:

$$A_1 \sin(\omega t + \varphi_1) + A_2 \sin(\omega t + \varphi_2) = A \sin(\omega t + \varphi) \tag{2.137a}$$

with

$$A = \sqrt{{A_1}^2 + {A_2}^2 + 2A_1A_2 \cos(\varphi_2 - \varphi_1)}, \tag{2.137b}$$

$$\tan \varphi = \frac{A_1 \sin \varphi_1 + A_2 \sin \varphi_2}{A_1 \cos \varphi_1 + A_2 \cos \varphi_2}, \tag{2.137c}$$

where the quantities A and φ can be determined by a vector diagram **(Fig. 2.41a)**.

A *linear combination* of several sine functions with the same frequency is also possible and yields a general sine function (harmonic oscillation) with the same frequency:

$$\sum_i c_i A_i \sin(\omega t + \varphi_i) = A \sin(\omega t + \varphi). \tag{2.138}$$

2.7.3.3 Vector Diagram for Oscillations

The general sine function (2.135, 2.136) can be represented easily by the polar coordinates $\rho = A, \varphi$ and by the Cartesian coordinates $x = a$, $y = b$ (see 3.5.2.1, p. 190) in a plane. The sum of two such quantities then behaves as the sum of two summand vectors **(Fig. 2.41a)**. Similarly the sum of several vectors results in a *linear combination* of several general sine functions. This representation is called a *vector diagram*.

The quantity u for a given time t can be determined from the vector diagram with the help of **Fig. 2.41b**: First the time axis $OP(t)$ has to be put through the origin O, which rotates clockwise around O by a constant angular velocity ω. At start $t = 0$ the axes y and t coincide. Then at any time t the projection ON of the vector $\vec{\mathbf{u}}$ onto the time axis is equal to the absolute value of the general sine function $u = A\sin(\omega t + \varphi)$. For time $t = 0$ the value $u_0 = A\sin\varphi$ is the projection onto the y-axis **(Fig. 2.41b)**.

2.7.3.4 Damping of Oscillations

The function $u(t) = Ae^{-at}\sin(\omega t + \varphi_0) \quad (a,\, t > 0)$ (2.139)

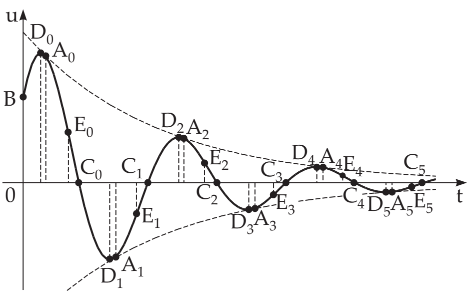


Figure 2.42

yields the *curve of a damped oscillation* **(Fig. 2.42)**.
The oscillation proceeds along the t-axis, while the curve asymptotically approaches the t-axis. The sine curve is enclosed by the exponential curves $u(t) = \pm Ae^{-at}$, and it contacts them in the points

$$A_0, A_1, A_2, \ldots, A_k = \left(\frac{\left(k + \frac{1}{2}\right)\pi - \varphi_0}{\omega}, \right.$$

$$\left. (-1)^k A \exp\left(-a \frac{\left(k + \frac{1}{2}\right)\pi - \varphi_0}{\omega} \right) \right).$$

The intersection points with the coordinate axes are $B = (0, A\sin\varphi_0), C_0, C_1, C_2, \ldots, C_k = \left(\frac{k\pi - \varphi_0}{\omega}, 0 \right)$. The extrema $D_0, D_1, D_2, \ldots$ are at $t_k = \frac{k\pi - \varphi_0 + \alpha}{\omega}$; and the inflection points $E_0, E_1, E_2, \ldots$ are at $t_k = \frac{k\pi - \varphi_0 + 2\alpha}{\omega}$ with $\tan\alpha = \frac{\omega}{a}$.

The *logarithmic decrement* of the damping is $\delta = \ln\left|\frac{y_i}{y_{i+1}}\right| = a\frac{\pi}{\omega}$, where y_i and y_{i+1} are the ordinates of two consecutive extrema.

2.8 Cyclometric or Inverse Trigonometric Functions

The *cyclometric functions* or *arcus functions* are the inverses of the trigonometric functions. For a unequivocal definition the domain of the trigonometric functions is to be decomposed into monotony intervals, to get an inverse function for every monotony interval. So, there are infinitely many such intervals, and for each its inverse is to be defined. In order to distinguish them an index k is to be assigned according to the corresponding interval. Obviously the trigonometric inverse functions are monotony in these intervals.

2.8.1 Definition of the Inverse Trigonometric Functions

How to define the inverse trigonometric functions will be shown here for the inverse of the sin function **(Fig. 2.43)**. The usual notation for it is $\arcsin x$. The domain of $y = \sin x$ will be split into monotony

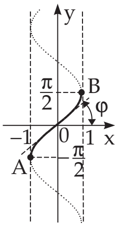


Figure 2.43

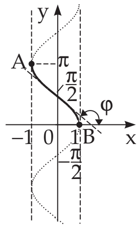


Figure 2.44

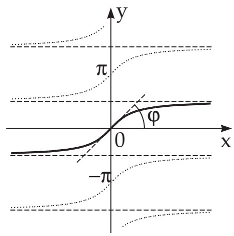


Figure 2.45

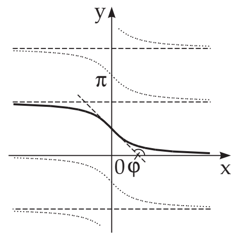


Figure 2.46

intervals $k\pi - \frac{\pi}{2} \leq x \leq k\pi + \frac{\pi}{2}$ with $k = 0, \pm 1, \pm 2, \ldots$. Reflecting the curve of $y = \sin x$ in the line $y = x$ yields the curve of the inverse function

$$y = \mathrm{arc}_k \sin x \tag{2.140a}$$

with the domains and ranges

$$-1 \leq x \leq +1 \quad \text{and} \quad k\pi - \frac{\pi}{2} \leq y \leq k\pi + \frac{\pi}{2}, \quad \text{where} \quad k = 0, \pm 1, \pm 2, \ldots . \tag{2.140b}$$

The form $y = \mathrm{arc}_k \sin x$ has the same meaning as $x = \sin y$.

Similarly, one can get the other inverse trigonometric functions which are represented in **Fig. 2.44–2.46**. The domains and ranges of the inverse functions can be found in **Table 2.6**.

2.8.2 Reduction to the Principal Value

In their domain the arcus functions have the so-called *principal values* for $k = 0$, written usually without an index, g.e., as $\arcsin x \equiv \mathrm{arc}_0 \sin x$. In **Fig. 2.47** the principal values of the inverse functions are presented.

The values of the different inverses can be calculated from the principal values by the following formulas:

$$\mathrm{arc}_k \sin x = k\pi + (-1)^k \arcsin x. \tag{2.141}$$

$$\mathrm{arc}_k \cos x = \begin{cases} (k+1)\pi - \arccos x & (k \text{ odd}), \\ k\pi + \arccos x & (k \text{ even}). \end{cases} \tag{2.142}$$

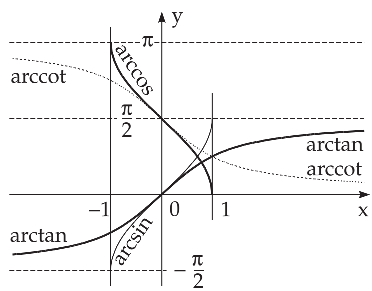


Figure 2.47

$$\mathrm{arc}_k \tan x = k\pi + \arctan x. \tag{2.143}$$

$$\mathrm{arc}_k \cot x = k\pi + \mathrm{arccot}\, x. \tag{2.144}$$

■ **A:** $\arcsin 0 = 0\,,\ \mathrm{arc}_k \sin 0 = k\pi.$

■ **B:** $\mathrm{arccot}\, 1 = \frac{\pi}{4}\,,\ \mathrm{arc}_k \cot 1 = \frac{\pi}{4} + k\pi.$

■ **C:** $\arccos \frac{1}{2} = \frac{\pi}{3},\ \mathrm{arc}_k \cos \frac{1}{2} = -\frac{\pi}{3} + (k+1)\pi$ for odd k,

$= \frac{\pi}{3} + k\pi$ for even k.

Remark: Calculators give the principal values of the trigonometric inverses.

Table 2.6 Domains and ranges of the inverses of trigonometric functions

Inverse function	Domain	Range	Trigonometric function with the same meaning
arc sine $y = \operatorname{arc}_k \sin x$	$-1 \le x \le 1$	$k\pi - \frac{\pi}{2} \le y \le k\pi + \frac{\pi}{2}$	$x = \sin y$
arc cosine $y = \operatorname{arc}_k \cos x$	$-1 \le x \le 1$	$k\pi \le y \le (k+1)\pi$	$x = \cos y$
arc tangent $y = \operatorname{arc}_k \tan x$	$-\infty < x < \infty$	$k\pi - \frac{\pi}{2} < y < k\pi + \frac{\pi}{2}$	$x = \tan y$
arc cotangent $y = \operatorname{arc}_k \cot x$	$-\infty < x < \infty$	$k\pi < y < (k+1)\pi$	$x = \cot y$

$k = 0, \pm 1, \pm 2, \ldots$. For $k = 0$ one gets the principal value of the inverse functions, which is usually written without an index, e.g., $\arcsin x \equiv \operatorname{arc}_0 \sin x$.

2.8.3 Relations Between the Principal Values

$$\arcsin x = \frac{\pi}{2} - \arccos x = \arctan \frac{x}{\sqrt{1-x^2}} = \begin{cases} -\arccos \sqrt{1-x^2} & (-1 \le x \le 0), \\ \arccos \sqrt{1-x^2} & (0 \le x \le 1). \end{cases} \tag{2.145}$$

$$\arccos x = \frac{\pi}{2} - \arcsin x = \operatorname{arccot} \frac{x}{\sqrt{1-x^2}} = \begin{cases} \pi - \arcsin \sqrt{1-x^2} & (\pi - 1 \le x \le 0), \\ \arcsin \sqrt{1-x^2} & (0 \le x \le 1). \end{cases} \tag{2.146}$$

$$\arctan x = \frac{\pi}{2} - \operatorname{arccot} x = \arcsin \frac{x}{\sqrt{1+x^2}} . \tag{2.147}$$

$$\arctan x = \begin{cases} \operatorname{arccot} \frac{1}{x} - \pi & (x < 0) \\ \operatorname{arccot} \frac{1}{x} & (x > 0) \end{cases} = \begin{cases} -\arccos \frac{1}{\sqrt{1+x^2}} & (x \le 0), \\ \arccos \frac{1}{\sqrt{1+x^2}} & (x \ge 0). \end{cases} \tag{2.148}$$

$$\operatorname{arccot} x = \frac{\pi}{2} - \arctan x = \arccos \frac{x}{\sqrt{1+x^2}} . \tag{2.149}$$

$$\operatorname{arccot} x = \begin{cases} \arctan \frac{1}{x} + \pi & (x < 0) \\ \arctan \frac{1}{x} & (x > 0) \end{cases} = \begin{cases} \pi - \arcsin \frac{1}{\sqrt{1+x^2}} & (x \le 0), \\ \arcsin \frac{1}{\sqrt{1+x^2}} & (x \ge 0). \end{cases} \tag{2.150}$$

2.8.4 Formulas for Negative Arguments

$$\arcsin(-x) = -\arcsin x. \tag{2.151}$$

$$\arctan(-x) = -\arctan x. \tag{2.152}$$

$$\arccos(-x) = \pi - \arccos x. \tag{2.153}$$

$$\operatorname{arccot}(-x) = \pi - \operatorname{arccot} x. \tag{2.154}$$

2.8.5 Sum and Difference of arcsin x and arcsin y

$$\arcsin x + \arcsin y = \arcsin\left(x\sqrt{1-y^2} + y\sqrt{1-x^2}\right) \quad (xy \le 0 \text{ or } x^2+y^2 \le 1), \tag{2.155a}$$

$$= \pi - \arcsin\left(x\sqrt{1-y^2} + y\sqrt{1-x^2}\right) \quad (x>0, y>0,\ x^2+y^2>1), \tag{2.155b}$$

$$= -\pi - \arcsin\left(x\sqrt{1-y^2} + y\sqrt{1-x^2}\right) \quad (x<0, y<0,\ x^2+y^2>1). \tag{2.155c}$$

$$\arcsin x - \arcsin y = \arcsin\left(x\sqrt{1-y^2} - y\sqrt{1-x^2}\right) \quad (xy \ge 0 \text{ or } x^2+y^2 \le 1), \tag{2.156a}$$

$$= \pi - \arcsin\left(x\sqrt{1-y^2} - y\sqrt{1-x^2}\right) \quad (x>0,\ y<0,\ x^2+y^2>1), \tag{2.156b}$$

$$= -\pi - \arcsin\left(x\sqrt{1-y^2} - y\sqrt{1-x^2}\right) \quad (x<0,\ y>0,\ x^2+y^2>1). \tag{2.156c}$$

2.8.6 Sum and Difference of arccos x and arccos y

$$\arccos x + \arccos y = \arccos\left(xy - \sqrt{1-x^2}\sqrt{1-y^2}\right) \quad (x+y \ge 0), \tag{2.157a}$$

$$= 2\pi - \arccos\left(xy - \sqrt{1-x^2}\sqrt{1-y^2}\right) \quad (x+y<0). \tag{2.157b}$$

$$\arccos x - \arccos y = -\arccos\left(xy + \sqrt{1-x^2}\sqrt{1-y^2}\right) \quad (x \ge y), \tag{2.158a}$$

$$= \arccos\left(xy + \sqrt{1-x^2}\sqrt{1-y^2}\right) \quad (x<y). \tag{2.158b}$$

2.8.7 Sum and Difference of arctan x and arctan y

$$\arctan x + \arctan y = \arctan\frac{x+y}{1-xy} \quad (xy<1), \tag{2.159a}$$

$$= \pi + \arctan\frac{x+y}{1-xy} \quad (x>0,\ xy>1), \tag{2.159b}$$

$$= -\pi + \arctan\frac{x+y}{1-xy} \quad (x<0,\ xy>1). \tag{2.159c}$$

$$\arctan x - \arctan y = \arctan\frac{x-y}{1+xy} \quad (xy>-1), \tag{2.160a}$$

$$= \pi + \arctan\frac{x-y}{1+xy} \quad (x>0,\ xy<-1), \tag{2.160b}$$

$$= -\pi + \arctan\frac{x-y}{1+xy} \quad (x<0,\ xy<-1). \tag{2.160c}$$

2.8.8 Special Relations for arcsin x, arccos x, arctan x

$$2\arcsin x = \arcsin\left(2x\sqrt{1-x^2}\right) \quad \left(|x| \le \frac{1}{\sqrt{2}}\right), \tag{2.161a}$$

$$= \pi - \arcsin\left(2x\sqrt{1-x^2}\right) \quad \left(\frac{1}{\sqrt{2}} < x \le 1\right), \tag{2.161b}$$

$$= -\pi - \arcsin\left(2x\sqrt{1-x^2}\right) \quad \left(-1 \le x < -\frac{1}{\sqrt{2}}\right). \tag{2.161c}$$

$$2\arccos x = \arccos(2x^2-1) \quad (0 \le x \le 1), \tag{2.162a}$$

$$= 2\pi - \arccos(2x^2-1) \quad (-1 \le x < 0). \tag{2.162b}$$

$$2\arctan x = \arctan\frac{2x}{1-x^2} \quad (|x| < 1), \tag{2.163a}$$

$$= \pi + \arctan\frac{2x}{1-x^2} \quad (x > 1), \tag{2.163b}$$

$$= -\pi + \arctan\frac{2x}{1-x^2} \quad (x < -1). \tag{2.163c}$$

$$\cos(n\arccos x) = T_n(x) \quad (n \ge 1), \tag{2.164}$$

where $n \ge 1$ can also be a fractional number and $T_n(x)$ is given by the equation

$$T_n(x) = \frac{\left(x+\sqrt{x^2-1}\right)^n + \left(x-\sqrt{x^2-1}\right)^n}{2}. \tag{2.165}$$

For any integer n, $T_n(x)$ is a polynomial of x (a *Chebyshev polynomial*). To study the properties of the Chebyshev polynomials see 19.6.3, p. 988.

2.9 Hyperbolic Functions

2.9.1 Definition of Hyperbolic Functions

Hyperbolic sine, *hyperbolic cosine* and *hyperbolic tangent* are defined by the following formulas:

$$\sinh x = \frac{e^x - e^{-x}}{2}, \tag{2.166}$$

$$\cosh x = \frac{e^x + e^{-x}}{2}, \tag{2.167}$$

$$\tanh x = \frac{e^x - e^{-x}}{e^x + e^{-x}}. \tag{2.168}$$

The geometric definition (see 3.1.2.2, p. 132), is an analogy to the trigonometric functions.
Hyperbolic cotangent, *hyperbolic secant* and *hyperbolic cosecant* are defined as reciprocal values of the above hyperbolic functions:

$$\coth x = \frac{1}{\tanh x} = \frac{e^x + e^{-x}}{e^x - e^{-x}}, \tag{2.169}$$

$$\operatorname{sech} x = \frac{1}{\cosh x} = \frac{2}{e^x + e^{-x}}, \tag{2.170}$$

$$\operatorname{cosech} x = \frac{1}{\sinh x} = \frac{2}{e^x - e^{-x}}. \tag{2.171}$$

The shapes of curves of hyperbolic functions are shown in **Fig. 2.48–2.52**.

2.9.2 Graphical Representation of the Hyperbolic Functions

2.9.2.1 Hyperbolic Sine

$y = \sinh x$ (2.166) is an odd strictly monotone increasing function between $-\infty$ and $+\infty$ **(Fig. 2.49)**. The origin is its symmetry center, the inflection point, and here the angle of slope of the tangent line is $\varphi = \frac{\pi}{4}$. There is no asymptote.

2.9.2.2 Hyperbolic Cosine

$y = \cosh x$ (2.167) is an even function, it is strictly monotone decreasing for $x < 0$ from $+\infty$ to 1, and for $x > 0$ it is strictly monotone increasing from 1 until $+\infty$ **(Fig. 2.50)**. The minimum is at $x = 0$

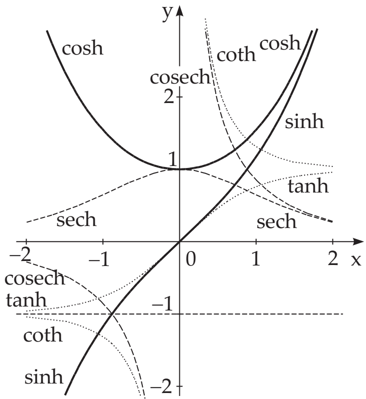


Figure 2.48

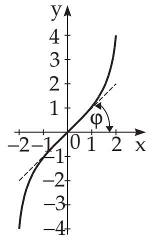


Figure 2.49

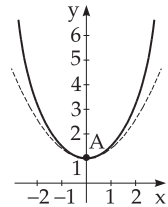


Figure 2.50

and it is equal to 1 (point $A(0,1)$); it has no asymptote. The curve is symmetric with respect to the y-axis and it always stays above the curve of the quadratic parabola $y = 1 + \frac{x^2}{2}$ (the broken-line curve). Because the function demonstrates a *catenary curve*, the curve is called the *catenoid* (see 2.15.1, p. 107).

2.9.2.3 Hyperbolic Tangent

$y = \tanh x$ (2.168) is an odd function, for $-\infty < x < +\infty$ strictly monotone increasing from -1 to $+1$ **(Fig. 2.51)**. The origin is the center of symmetry, and the inflection point, and here the angle of slope of the tangent line is $\varphi = \frac{\pi}{4}$. The asymptotes are the lines $y = \pm 1$.

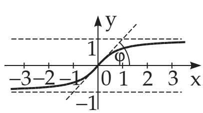


Figure 2.51

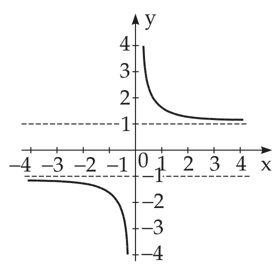


Figure 2.52

2.9.2.4 Hyperbolic Cotangent

$y = \coth x$ (2.169) is an odd function which is not continuous at $x = 0$ **(Fig. 2.52)**. It is strictly monotone decreasing in the interval $-\infty < x < 0$ and it takes its values from -1 until $-\infty$; in the interval $0 < x < +\infty$ it is also strictly monotone decreasing with values from $+\infty$ to $+1$. It has no

inflection point, no extreme value. The asymptotes are the lines $x = 0$ and $y = \pm 1$.

2.9.3 Important Formulas for the Hyperbolic Functions

There are similar relations between the hyperbolic functions as between trigonometric functions. The validity of the following formulas can be shown directly from the definitions of hyperbolic functions, or considering the definitions and relations of these functions also for complex arguments, from (2.199)–(2.206), they can be calculated from the formulas known for trigonometric functions.

2.9.3.1 Hyperbolic Functions of One Variable

$$\cosh^2 x - \sinh^2 x = 1, \quad (2.172) \qquad \coth^2 x - \operatorname{cosech}^2 x = 1, \quad (2.173)$$

$$\operatorname{sech}^2 x + \tanh^2 x = 1, \quad (2.174) \qquad \tanh x \cdot \coth x = 1, \quad (2.175)$$

$$\frac{\sinh x}{\cosh x} = \tanh x, \quad (2.176) \qquad \frac{\cosh x}{\sinh x} = \coth x. \quad (2.177)$$

2.9.3.2 Expressing a Hyperbolic Function by Another One with the Same Argument

The corresponding formulas are collected in **Table 2.7**.

2.9.3.3 Formulas for Negative Arguments

$$\sinh(-x) = -\sinh x, \quad (2.178) \qquad \cosh(-x) = \cosh x, \quad (2.180)$$

$$\tanh(-x) = -\tanh x, \quad (2.179) \qquad \coth(-x) = -\coth x. \quad (2.181)$$

Table 2.7 Relations between two hyperbolic functions with the same arguments for $x > 0$

	$\sinh x$	$\cosh x$	$\tanh x$	$\coth x$
$\sinh x$	–	$\sqrt{\cosh^2 x - 1}$	$\dfrac{\tanh x}{\sqrt{1 - \tanh^2 x}}$	$\dfrac{1}{\sqrt{\coth^2 x - 1}}$
$\cosh x$	$\sqrt{\sinh^2 x + 1}$	–	$\dfrac{1}{\sqrt{1 - \tanh^2 x}}$	$\dfrac{\coth x}{\sqrt{\coth^2 x - 1}}$
$\tanh x$	$\dfrac{\sinh x}{\sqrt{\sinh^2 x + 1}}$	$\dfrac{\sqrt{\cosh^2 x - 1}}{\cosh x}$	–	$\dfrac{1}{\coth x}$
$\coth x$	$\dfrac{\sqrt{\sinh^2 x + 1}}{\sinh x}$	$\dfrac{\cosh x}{\sqrt{\cosh^2 x - 1}}$	$\dfrac{1}{\tanh x}$	–

2.9.3.4 Hyperbolic Functions of the Sum and Difference of Two Arguments (Addition Theorems)

$$\sinh(x \pm y) = \sinh x \cosh y \pm \cosh x \sinh y, \quad (2.182)$$

$$\cosh(x \pm y) = \cosh x \cosh y \pm \sinh x \sinh y, \quad (2.183)$$

$$\tanh(x \pm y) = \frac{\tanh x \pm \tanh y}{1 \pm \tanh x \tanh y}, \quad (2.184) \qquad \coth(x \pm y) = \frac{1 \pm \coth x \coth y}{\coth x \pm \coth y}. \quad (2.185)$$

2.9.3.5 Hyperbolic Functions of Double Arguments

$$\sinh 2x = 2 \sinh x \cosh x, \quad (2.186)$$

$$\cosh 2x = \sinh^2 x + \cosh^2 x, \quad (2.187)$$

$$\tanh 2x = \frac{2 \tanh x}{1 + \tanh^2 x}, \quad (2.188)$$

$$\coth 2x = \frac{1 + \coth^2 x}{2 \coth x}. \quad (2.189)$$

2.9.3.6 De Moivre Formula for Hyperbolic Functions

$$(\cosh x \pm \sinh x)^n = \left(e^{\pm x}\right)^n = e^{\pm nx} = \cosh nx \pm \sinh nx. \quad (2.190)$$

2.9.3.7 Hyperbolic Functions of Half-Argument

$$\sinh \frac{x}{2} = \pm\sqrt{\frac{1}{2}(\cosh x - 1)}, \quad (2.191) \qquad \cosh \frac{x}{2} = \sqrt{\frac{1}{2}(\cosh x + 1)}, \quad (2.192)$$

The sign of the square root in (2.191) is positive for $x > 0$ and negative for $x < 0$.

$$\tanh \frac{x}{2} = \frac{\cosh x - 1}{\sinh x} = \frac{\sinh x}{\cosh x + 1}, \quad (2.193) \qquad \coth \frac{x}{2} = \frac{\sinh x}{\cosh x - 1} = \frac{\cosh x + 1}{\sinh x}. \quad (2.194)$$

2.9.3.8 Sum and Difference of Hyperbolic Functions

$$\sinh x \pm \sinh y = 2 \sinh \frac{x \pm y}{2} \cosh \frac{x \mp y}{2}, \quad (2.195)$$

$$\cosh x + \cosh y = 2 \cosh \frac{x + y}{2} \cosh \frac{x - y}{2}, \quad (2.196)$$

$$\cosh x - \cosh y = 2 \sinh \frac{x + y}{2} \sinh \frac{x - y}{2}, \quad (2.197)$$

$$\tanh x \pm \tanh y = \frac{\sinh(x \pm y)}{\cosh x \cosh y}. \quad (2.198)$$

2.9.3.9 Relation Between Hyperbolic and Trigonometric Functions with Complex Arguments z

$$\sin z = -\mathrm{i} \sinh \mathrm{i}z, \quad (2.199) \qquad \sinh z = -\mathrm{i} \sin \mathrm{i}z, \quad (2.203)$$

$$\cos z = \cosh \mathrm{i}z, \quad (2.200) \qquad \cosh z = \cos \mathrm{i}z, \quad (2.204)$$

$$\tan z = -\mathrm{i} \tanh \mathrm{i}z, \quad (2.201) \qquad \tanh z = -\mathrm{i} \tan \mathrm{i}z, \quad (2.205)$$

$$\cot z = \mathrm{i} \coth \mathrm{i}z, \quad (2.202) \qquad \coth z = \mathrm{i} \cot \mathrm{i}z. \quad (2.206)$$

Every relation between hyperbolic functions, which contains x or ax but not $ax+b$, can be derived from the corresponding trigonometric relation with the substitution $\mathrm{i}\sinh x$ for $\sin \alpha$ and $\cosh x$ for $\cos \alpha$.

■ **A:** $\cos^2 \alpha + \sin^2 \alpha = 1$, $\cosh^2 x + \mathrm{i}^2 \sinh^2 x = 1$ or $\cosh^2 x - \sinh^2 x = 1$.

■ **B:** $\sin 2\alpha = 2\sin\alpha\cos\alpha$, $\mathrm{i}\sinh 2x = 2\mathrm{i}\sinh x\cosh x$ or $\sinh 2x = 2\sinh x\cosh x$.

2.10 Area Functions

2.10.1 Definitions

The *area functions* are the inverse functions of the hyperbolic functions, i.e., the *inverse hyperbolic functions*. The functions $\sinh x$, $\tanh x$, and $\coth x$ are strictly monotone, so they have unique inverses without any restriction; the function $\cosh x$ has two monotonic intervals so there are to consider two inverse functions. The name *area* refers to the fact that the geometric definition of the functions is the area of certain hyperbolic sectors (see 3.1.2.2, p. 132).

2.10.1.1 Area Sine

The function $y = \operatorname{Arsinh} x$ (2.207)

(Fig. 2.53) is an odd, strictly monotone increasing function, with domain and range given in **Table 2.8**. (2.53) is equivalent to the expression $x = \sinh y$. The origin is the center of symmetry and the inflection point of the curve, where the angle of slope of the tangent line is $\varphi = \dfrac{\pi}{4}$.

2.10.1.2 Area Cosine

The functions $y = \operatorname{Arcosh} x$ and $y = -\operatorname{Arcosh} x$ (2.208)

(Fig. 2.54) or $x = \cosh y$ have the domain and range given in **Table 2.8**; they are defined only for $x \geq 1$. The function curve starts at the point $A(1, 0)$ with a vertical tangent line and the function increases or decreases strictly monotonically respectively.

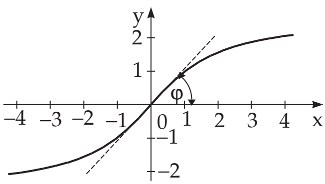


Figure 2.53

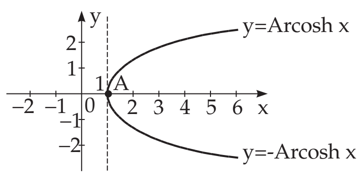


Figure 2.54

Table 2.8 Domains and ranges of the area functions

Area function	Domain	Range	Hyperbolic function with same meaning
area sine $y = \operatorname{Arsinh} x$	$-\infty < x < \infty$	$-\infty < y < \infty$	$x = \sinh y$
area cosine $y = \operatorname{Arcosh} x$ $y = -\operatorname{Arcosh} x$	$1 \leq x < \infty$	$0 \leq y < \infty$ $-\infty < y \leq 0$	$x = \cosh y$
area tangent $y = \operatorname{Artanh} x$	$\lvert x\rvert < 1$	$-\infty < y < \infty$	$x = \tanh y$
area cotangent $y = \operatorname{Arcoth} x$	$\lvert x\rvert > 1$	$-\infty < y < 0$ $0 < y < \infty$	$x = \coth y$

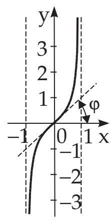


Figure 2.55

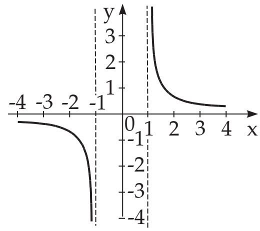


Figure 2.56

2.10.1.3 Area Tangent

The function $\quad y = \operatorname{Artanh} x$ (2.209)

(Fig. 2.55) or $x = \tanh y$ is an odd function, defined only for $|x| < 1$, with domain and range given in **Table 2.8**. The origin is the center of symmetry and also the inflection point of the curve, and here the angle of slope of the tangent line is $\varphi = \dfrac{\pi}{4}$. The asymptotes are vertical, their equations are $x = \pm 1$.

2.10.1.4 Area Cotangent

The function $\quad y = \operatorname{Arcoth} x$ (2.210)

(Fig. 2.56) or $x = \coth y$ is an odd function, defined only for $|x| > 1$, with domain and range given in **Table 2.8**. In the interval $-\infty < x < -1$ the function is strictly monotone decreasing from 0 until $-\infty$, in the interval $1 < x < +\infty$ it is strictly monotone decreasing from $+\infty$ to 0. It has three asymptotes: their equations are $y = 0$ and $x = \pm 1$.

2.10.2 Determination of Area Functions Using Natural Logarithm

From the definition of hyperbolic functions ((2.166)–(2.171), see 2.9.1, p. 89) follows that the area functions can be expressed with the logarithm function:

$$\operatorname{Arsinh} x = \ln\left(x + \sqrt{x^2+1}\right), \tag{2.211}$$

$$\operatorname{Arcosh} x = \ln\left(x + \sqrt{x^2-1}\right) = \ln\left(\frac{1}{x - \sqrt{x^2-1}}\right) \quad (x \geq 1), \tag{2.212}$$

$$\operatorname{Artanh} x = \frac{1}{2}\ln\frac{1+x}{1-x} \quad (|x| < 1), \tag{2.213}$$

$$\operatorname{Arcoth} x = \frac{1}{2}\ln\frac{x+1}{x-1} \quad (|x| > 1). \tag{2.214}$$

2.10.3 Relations Between Different Area Functions

$$\operatorname{Arsinh} x = (\operatorname{sign} x)\operatorname{Arcosh}\sqrt{x^2+1} = \operatorname{Artanh}\frac{x}{\sqrt{x^2+1}} = \operatorname{Arcoth}\frac{\sqrt{x^2+1}}{x} \quad (|x| < \infty), \tag{2.215}$$

$$\operatorname{Arcosh} x = \operatorname{Arsinh}\sqrt{x^2-1} = \operatorname{Artanh}\frac{\sqrt{x^2-1}}{x} = \operatorname{Arcoth}\frac{x}{\sqrt{x^2-1}} \quad (x \geq 1), \tag{2.216}$$

$$\operatorname{Artanh} x = \operatorname{Arsinh}\frac{x}{\sqrt{1-x^2}} = \operatorname{Arcoth}\frac{1}{x} = (\operatorname{sign} x)\operatorname{Arcosh}\frac{1}{\sqrt{1-x^2}} \quad (|x| < 1), \tag{2.217}$$

$$\operatorname{Arcoth} x = \operatorname{Artanh}\frac{1}{x} = (\operatorname{sign} x)\operatorname{Arsinh}\frac{1}{\sqrt{x^2-1}}$$
$$= (\operatorname{sign} x)\operatorname{Arcosh}\frac{|x|}{\sqrt{x^2-1}} \qquad (|x|>1). \tag{2.218}$$

2.10.4 Sum and Difference of Area Functions

$$\operatorname{Arsinh} x \pm \operatorname{Arsinh} y = \operatorname{Arsinh}\left(x\sqrt{1+y^2} \pm y\sqrt{1+x^2}\right), \tag{2.219}$$

$$\operatorname{Arcosh} x \pm \operatorname{Arcosh} y = \operatorname{Arcosh}\left(xy \pm \sqrt{(x^2-1)(y^2-1)}\right), \tag{2.220}$$

$$\operatorname{Artanh} x \pm \operatorname{Artanh} y = \operatorname{Artanh}\frac{x \pm y}{1 \pm xy}. \tag{2.221}$$

2.10.5 Formulas for Negative Arguments

$$\operatorname{Arsinh}(-x) = -\operatorname{Arsinh} x, \tag{2.222}$$

$$\operatorname{Artanh}(-x) = -\operatorname{Artanh} x, \quad (2.223) \qquad \operatorname{Arcoth}(-x) = -\operatorname{Arcoth} x. \quad (2.224)$$

The functions Arsinh, Artanh and Arcoth are odd functions, and Arcosh (2.212) is not defined for arguments $x < 1$.

2.11 Curves of Order Three (Cubic Curves)

A curve is called an algebraic curve of order n if it can be written in the form of a polynomical equation $F(x, y) = 0$ of two variables where the left-hand side is a poynomial expression of degree n.

■ The cardioid with equation $(x^2+y^2)(x^2+y^2-2ax)-a^2y^2 = 0$ $(a > 0)$ (see 2.12.2, p. 98) is a curve of order four. The well-known conic sections (see 3.5.2.11, p. 206) result in curves of order two.

2.11.1 Semicubic Parabola

The equation $\quad y = ax^{3/2} \quad (a > 0,\ x \ge 0)$ (2.225a)

or in parametric form $\quad x = t^2, \quad y = at^3 \quad (a > 0,\ -\infty < t < \infty)$ (2.225b)

gives the *semicubic parabola* **(Fig. 2.57)**. It has a cuspidal point at the origin, it has no asymptote. The curvature $K = \dfrac{6a}{\sqrt{x}(4+9a^2x)^{3/2}}$ takes all the values between ∞ and 0. The arclength of the curve between the origin and the point $P(x, y)$ is $L = \dfrac{1}{27a^2}[(4+9a^2x)^{3/2} - 8]$.

2.11.2 Witch of Agnesi

The equation $\qquad y = \dfrac{a^3}{a^2+x^2} \quad (a > 0,\ -\infty < x < \infty)$ (2.226a)

determines the curve represented in **Fig. 2.58**, the *witch of Agnesi*. It has an asymptote with the equation $y = 0$, it has an extreme point at $A(0, a)$, where the radius of curvature is $r = \dfrac{a}{2}$. The inflection points B and C are at $\left(\pm\dfrac{a}{\sqrt{3}}, \dfrac{3a}{4}\right)$, where the angles of slope of the tangent lines are $\tan\varphi = \mp\dfrac{3\sqrt{3}}{8}$.

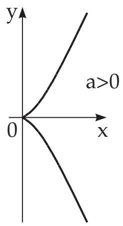


Figure 2.57

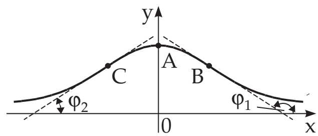


Figure 2.58

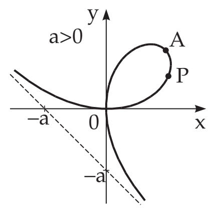


Figure 2.59

The area of the region between the curve and its asymptote is equal to $S = \pi a^2$. The witch of Agnesi (2.226a) is a special case of the Lorentz or Breit–Wigner curve

$$y = \frac{a}{b^2 + (x - c)^2} \quad (a > 0\,, b \neq 0)\,. \tag{2.226b}$$

■ The Fourier transform of the damped oscillation is the Lorentz or Breit–Wigner curve (see 15.3.1.4, p. 791).

2.11.3 Cartesian Folium (Folium of Descartes)

The equation $\quad x^3 + y^3 = 3axy \quad (a > 0) \quad$ or (2.227a)

in parametric form $\quad x = \frac{3at}{1+t^3}\,,\ y = \frac{3at^2}{1+t^3} \quad$ with

$$t = \tan \sphericalangle P0x \quad (a > 0, -\infty < t < -1 \ \text{ and } \ -1 < t < \infty) \tag{2.227b}$$

gives the *Cartesian folium curve* represented in **Fig. 2.59**. The origin is a double point because the curve passes through it twice, and here both coordinate axes are tangent lines. At the origin the radius of curvature for both branches of the curve is $r = \frac{3a}{2}$. The equation of the asymptote is $x + y + a = 0$. The vertex A has the coordinates $A\left(\frac{3}{2}a, \frac{3}{2}a\right)$. The area of the loop is $S_1 = \frac{3a^2}{2}$. The area S_2 between the curve and the asymptote has the same value.

2.11.4 Cissoid

The equation $\quad y^2 = \frac{x^3}{a - x} \quad (a > 0),$ (2.228a)

or in parametric form $\quad x = \frac{at^2}{1+t^2}\,,\ y = \frac{at^3}{1+t^2} \quad$ with

$$t = \tan \sphericalangle P0x \quad (a > 0, -\infty < t < \infty) \tag{2.228b}$$

or with polar coordinates $\rho = \frac{a \sin^2 \varphi}{\cos \varphi} \quad (a > 0)$ (2.228c)

(Fig. 2.60) describes the locus of the points P for which

$$\overline{0P} = \overline{MQ} \tag{2.229}$$

is valid. Here M is the second intersection point of the line $0P$ with the drawn circle of radius $\frac{a}{2}$, and Q is the intersection point of the line $0P$ with the asymptote $x = a$. The area between the curve and the asymptote is equal to $S = \frac{3}{4}\pi a^2$.

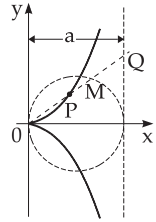


Figure 2.60

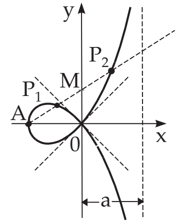


Figure 2.61

2.11.5 Strophoide

Strophoide is the locus of the points P_1 and P_2, which are on an arbitrary half-line starting at A (A is on the negative x-axis) and for which the equalities

$$\overline{MP_1} = \overline{MP_2} = \overline{0M} \tag{2.230}$$

are valid. Here M is the intersection point with the y-axis **(Fig. 2.61)**. The equation of the strophoide in Cartesian, and in polar coordinates, and in parametric form is:

$$y^2 = x^2\left(\frac{a+x}{a-x}\right) \quad (a > 0), \qquad (2.231\text{a}) \qquad\qquad \rho = -a\frac{\cos 2\varphi}{\cos\varphi} \quad (a > 0), \qquad (2.231\text{b})$$

$$x = a\frac{t^2-1}{t^2+1}, \quad y = at\frac{t^2-1}{t^2+1} \text{ with } t = \tan \sphericalangle P0x \quad (a > 0, -\infty < t < \infty). \tag{2.231c}$$

The origin is a double point with tangent lines $y = \pm x$. The asymptote has the equation $x = a$. The vertex is $A(-a, 0)$. The area of the loop is $S_1 = 2a^2 - \frac{1}{2}\pi a^2$, and the area between the curve and the asymptote is $S_2 = 2a^2 + \frac{1}{2}\pi a^2$.

2.12 Curves of Order Four (Quartics)

2.12.1 Conchoid of Nicomedes

The *Conchoid of Nicomedes* **(Fig. 2.62)** is the locus of the points P, for which

$$\overline{0P} = \overline{0M} \pm l \tag{2.232}$$

holds, where M is the intersection point of the line $\overline{0P_1 0P_2}$ with the asymptote $x = a$. The "+" sign belongs to the right branch of the curve, the "−" sign belongs to the left one in relation to the asymptote. The equations for the *conchoid of Nicomedes* are the following in Cartesian coordinates, in parametric form and in polar coordinates:

$$(x-a)^2(x^2+y^2) - l^2x^2 = 0 \quad (a > 0, l > 0), \tag{2.233a}$$

$$x = a + l\cos\varphi, \quad y = a\tan\varphi + l\sin\varphi$$
$$\left(a > 0, \text{ right branch: } -\frac{\pi}{2} < \varphi < \frac{\pi}{2}, \text{ left branch: } \frac{\pi}{2} < \varphi < \frac{3\pi}{2}\right), \tag{2.233b}$$

$$\rho = \frac{a}{\cos\varphi} \pm l \quad (\text{“+” sign: right branch, “−” sign: left branch,}) \tag{2.233c}$$

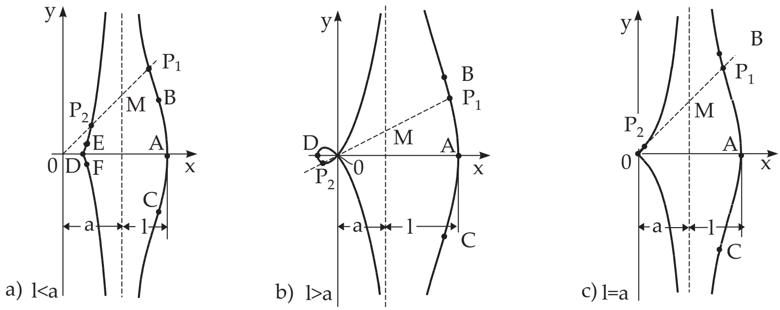


Figure 2.62

1. Right Branch: The asymptote is $x = a$. The vertex A is at $(a + l, 0)$, the inflection points B, C have as x-coordinate the greatest root of the equation $x^3 - 3a^2x + 2a(a^2 - l^2) = 0$. The area between the right branch and the asymptote is $S = \infty$.

2. Left Branch: The asymptote is $x = a$. The vertex D is at $(a - l, 0)$. The origin is a singular point, whose type depends on a and l:

Case a) For $l < a$ it is an isolated point **(Fig. 2.62a)**. The curve has two further inflection points E and F, whose abscissa is the second greatest root of the equation $x^3 - 3a^2x + 2a(a^2 - l^2) = 0$.

Case b) For $l > a$ the origin is a double point **(Fig. 2.62b)**. The curve has a maximum and a minimum value at $x = a - \sqrt[3]{al^2}$. At the origin the slopes of the tangent lines are $\tan\alpha = \dfrac{\pm\sqrt{l^2 - a^2}}{a}$. Here the radius of curvature is $r_0 = \dfrac{l\sqrt{l^2 - a^2}}{2a}$.

Case c) For $l = a$ the origin is a cuspidal point **(Fig. 2.62c)**.

2.12.2 General Conchoid

The *conchoid of Nicomedes* is a special case of the *general conchoid.* One gets the conchoid of a given curve by elongating the length of the position vector of every point by a given constant segment $\pm l$. Considering a curve in a polar coordinate system with an equation $\rho = f(\varphi)$, then the equation of its conchoid is

$$\rho = f(\varphi) \pm l. \tag{2.234}$$

So, the conchoid of Nicomedes is the *conchoid of the line.*

2.12.3 Pascal's Limaçon

The *conchoid of a circle* is called the *Pascal limaçon* **(Fig. 2.63)** if in (2.232) the origin is on the perimeter of the circle, which is a further special case of the general conchoid (see 2.12.2, p. 98). The equations in the Cartesian and in the polar coordinate systems and in parametric form are the following (see also (2.246c), p. 105):

$$(x^2 + y^2 - ax)^2 = l^2(x^2 + y^2) \quad (a > 0,\ l > 0), \tag{2.235a}$$

$$\rho = a\cos\varphi + l \quad (a > 0,\ l > 0), \tag{2.235b}$$

$$x = a\cos^2\varphi + l\cos\varphi, \quad y = a\cos\varphi\sin\varphi + l\sin\varphi \quad (a > 0,\ l > 0,\ 0 \le \varphi < 2\pi) \tag{2.235c}$$

with a as the diameter of the circle. The vertices A, B are at $(a \pm l, 0)$. The shape of the curve depends on the quantities a and l, as can be seen in **Fig. 2.63** and **Fig. 2.64**.

a) Extreme Points and Inflection Points: For $a > l$ the curve has four extreme points C, D, E, F; for $a \le l$ it has two; they are at $\left(\cos\varphi = \frac{-l \pm \sqrt{l^2 + 8a^2}}{4a}\right)$. For $a < l < 2a$ there exist two inflection points G and H at $\left(\cos\varphi = -\frac{2a^2 + l^2}{3al}\right)$.

b) Double Tangent: For $l < 2a$, at the points I and K at $\left(-\frac{l^2}{4a}, \pm\frac{l\sqrt{4a^2 - l^2}}{4a}\right)$ there is a double tangent.

c) Singular Points: The origin is a singular point: For $a < l$ it is an isolated point, for $a > l$ it is a

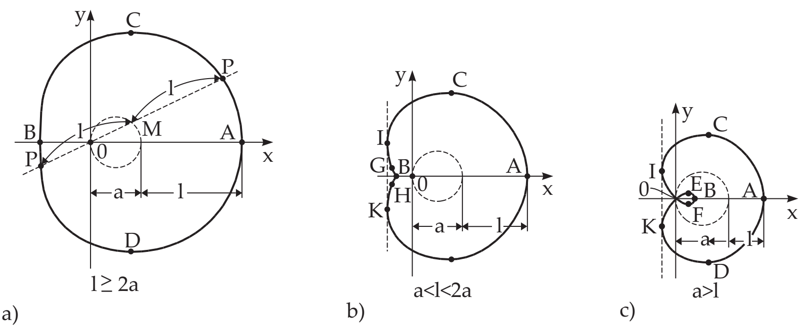


Figure 2.63

double point and the slopes of the tangent lines are $\tan\alpha = \pm\frac{\sqrt{a^2 - l^2}}{l}$, here the radius of curvature is $r_0 = \frac{1}{2}\sqrt{a^2 - l^2}$.

For $a = l$ the origin is a cuspidal point; then the curve is called a *cardioid* (see also 2.13.3, p. 103).

The area of the limaçon is $S = \frac{\pi a^2}{2} + \pi l^2$, where in the case $a > l$ **(Fig. 2.63c)** the area of the inside loop is counted twice.

2.12.4 Cardioid

The *cardioid* **(Fig. 2.64)** can be defined in two different ways, as:

1. Special case of the *Pascal limaçon* with

$$\overline{0P} = \overline{0M} \pm a, \tag{2.236}$$

where a is the diameter of the circle.

2. Special case of the *epicycloid* with the same diameter a for the fixed and for the moving circle (see 2.13.3, p. 103). The equation is

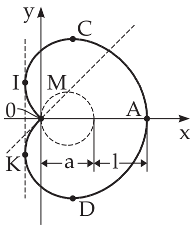


Figure 2.64

$$(x^2+y^2)^2 - 2ax(x^2+y^2) = a^2y^2 \quad (a>0), \tag{2.237a}$$

and the parametric form, and the equation in polar coordinates are:

$$x = a\cos\varphi(1+\cos\varphi), \quad y = a\sin\varphi(1+\cos\varphi)$$
$$(a>0, 0 \le \varphi < 2\pi), \tag{2.237b}$$
$$\rho = a(1+\cos\varphi) \quad (a>0). \tag{2.237c}$$

The origin is a cuspidal point. The vertex A is at $(2a,0)$; extreme points C and D are at $\cos\varphi = \frac{1}{2}$ with coordinates $\left(\frac{3}{4}a\,,\pm\frac{3\sqrt{3}}{4}a\right)$. The area is $S = \frac{3}{2}\pi a^2$, i.e., six times the area of a circle with diameter a. The length of the curve is $L = 8a$.

2.12.5 Cassinian Curve

The locus of the points P, for which the product of the distances from two fixed points F_1 and F_2 with coordinates $(c,0)$ and $(-c,0)$ resp., is equal to a constant $a^2 \neq 0$, is called a *Cassinian curve* **(Fig. 2.65)**:

$$\overline{F_1P}\cdot\overline{F_2P} = a^2. \tag{2.238}$$

The equations in Cartesian and polar coordinates are:

$$(x^2+y^2)^2 - 2c^2(x^2-y^2) = a^4 - c^4, \quad (a>0, c>0), \tag{2.239a}$$

$$\rho^2 = c^2\cos 2\varphi \pm \sqrt{c^4\cos^2 2\varphi + (a^4-c^4)} \quad (a>0, c>0). \tag{2.239b}$$

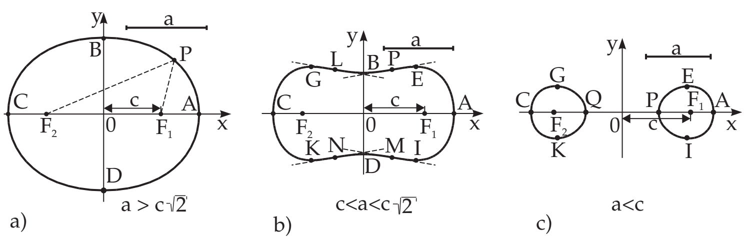


Figure 2.65

The shape of the curve depends on the quantities a and c :

Case $a > c\sqrt{2}$: For $a > c\sqrt{2}$ the curve is an oval whose shape resembles an ellipse **(Fig. 2.65a)**. The intersection points A, C with the x-axis are $(\pm\sqrt{a^2+c^2}\,,0)$, the intersection points B, D with the y-axis are $(0,\ \pm\sqrt{a^2-c^2})$.

Case $a = c\sqrt{2}$: For $a = c\sqrt{2}$ the curve is of the same type with A, C $(\pm c\sqrt{3}\,,\ 0)$ and B, D $(0, \pm c)$, where the curvature at the points B and D is equal to 0, i.e., there is a narrow contact with the lines $y = \pm c$.

Case $c < a < c\sqrt{2}$: For $c < a < c\sqrt{2}$ the curve is a pressed oval **(Fig. 2.65b)**. The intersection points with the axes are the same as in the case $a > c\sqrt{2}$, also the extreme points B, D, while there are further extreme points E, G, K, I at $\left(\pm\frac{\sqrt{4c^4-a^4}}{2c}\,,\ \pm\frac{a^2}{2c}\right)$ and there are four inflection points P, L,

M, N at $\left(\pm\sqrt{\frac{1}{2}(m-n)}\,,\ \pm\sqrt{\frac{1}{2}(m+n)}\right)$ with $n = \frac{a^4 - c^4}{3c^2}$ and $m = \sqrt{\frac{a^4 - c^4}{3}}$.

Case $a = c$: For $a = c$ there is the *lemniscate*.

Case $a < c$: For $a < c$ there are two ovals **(Fig. 2.65c)**. The intersection points A, C and P, Q with the x-axis are at $(\pm\sqrt{a^2 + c^2}\,,\ 0)$ and $(\pm\sqrt{c^2 - a^2}, 0)$. The extreme points E, G, K, I are at $\left(\pm\frac{\sqrt{4c^4 - a^4}}{2c},\ \pm\frac{a^2}{2c}\right)$. The radius of curvature is $r = \frac{2a^2\rho^3}{c^4 - a^4 + 3\rho^4}$, where ρ satisfies the polar coordinate representation.

2.12.6 Lemniscate

The *lemniscate* **(Fig. 2.66)** is the special case $a = c$ of the *Cassinian curve* satisfying the condition

$$\overline{F_1P} \cdot \overline{F_2P} = \left(\frac{\overline{F_1F_2}}{2}\right)^2, \qquad (2.240)$$

where the fixed points F_1, F_2 are at $(\pm a, 0)$. The equation in Cartesian coordinates is

$$(x^2 + y^2)^2 - 2a^2(x^2 - y^2) = 0 \quad (a > 0) \qquad (2.241a)$$

and in polar coordinates

$$\rho = a\sqrt{2\cos 2\varphi} \quad (a > 0). \qquad (2.241b)$$

The origin is a double point and an inflection point at the same time, where the tangents are $y = \pm x$.

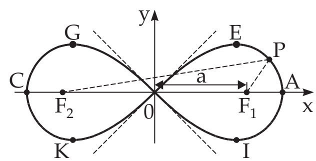


Figure 2.66

The intersection points A and C with the x-axis are at $(\pm a\sqrt{2}, 0)$, the extreme points of the curve E, G, K, I are at $\left(\pm\frac{a\sqrt{3}}{2},\ \pm\frac{a}{2}\right)$. The polar angle at these points is $\varphi = \pm\frac{\pi}{6}$. The radius of curvature is $r = \frac{2a^2}{3\rho}$ and the area of every loop is $S = a^2$.

2.13 Cycloids

2.13.1 Common (Standard) Cycloid

The *cycloid* is a curve which is described by a point of the perimeter of a circle while the circle rolls along a line without sliding **(Fig. 2.67)**. The equation of the *usual cycloid* written in parametric form is the following:

$$x = a(t - \sin t), \quad y = a(1 - \cos t)$$
$$(a > 0, \ -\infty < t < \infty), \qquad (2.242a)$$

where a is the radius of the circle and t is the angle $\measuredangle PC_1B$ in radian measure. In Cartesian coordinates

$$x + \sqrt{y(2a - y)} = a\operatorname{arc}_k \cos\frac{a - y}{a}$$
$$(a > 0, \ k = 0, \pm 1, \pm 2, \ldots) \qquad (2.242b)$$

holds.

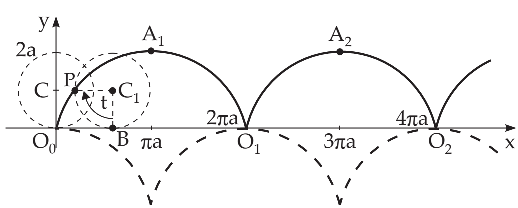


Figure 2.67

The curve is periodic with period $\overline{O_0O_1} = 2\pi a$. At $O_0, O_1, O_2, \ldots, O_k = (2k\pi a, 0)$ there are cusps, the vertices are at $A_{k+1} = ((2k+1)\pi a, 2a\ (k = 0, \pm 1, \pm 2, \ldots))$. The arc length of O_0P is $L = 8a\sin^2(t/4)$, the length of one arch is $L_{O_0A_1O_1} = 8a$. The area of one arch is $S = 3\pi a^2$. The radius of curvature is $r = 4a\sin\frac{1}{2}t$, at the vertices $r_A = 4a$. The evolute of a cycloid (see 3.6.1.6, p. 254) is a *congruent cycloid*, which is denoted in **Fig. 2.67** by the broken line.

2.13.2 Prolate and Curtate Cycloids or Trochoids

Prolate and *curtate cycloids* or *trochoids* are curves described by a point, which is inside or outside of a circle, fixed on a half-line starting from the center of the circle, while the circle rolls along a line without sliding **(Fig. 2.68).**
The equation of the trochoid in parametric form is

$$x = a(t - \lambda \sin t), \tag{2.243a}$$

$$y = a(1 - \lambda \cos t), \tag{2.243b}$$

where a is the radius of the circle, t is the angle $\sphericalangle PC_1M$, and $\lambda a = \overline{C_1P}$.
The case $\lambda > 1$ gives the prolate cycloid and $\lambda < 1$ the curtate one.
The period of the curve is $\overline{O_0O_1} = 2\pi a$, the maximum points are at $A_1, A_2, \ldots, A_{k+1} = ((2k+1)\pi a, (1+\lambda)a)$, the minimum points are at $B_0, B_1, B_2, \ldots, B_k = (2k\pi a, (1-\lambda)a)\ (k = 0, \pm 1, \pm 2, \ldots)$. The prolate cycloid has double points at $D_0, D_1, D_2, \ldots, D_k = \left[2k\pi a\, , a\left(1 - \sqrt{\lambda^2 - {t_0}^2}\right)\right]$, where t_0 is the smallest positive root of the equation $t = \lambda \sin t$.

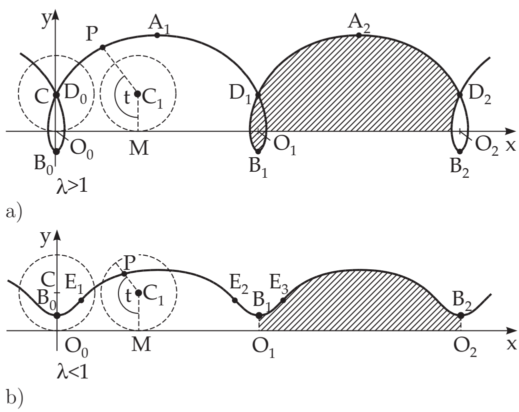


Figure 2.68

The curtate cycloid has inflection points at $E_1, E_2, \ldots, E_{k+1} = \left[a\left(\operatorname{arc}_k \cos\lambda - \lambda\sqrt{1-\lambda^2}\right), a\left(1-\lambda^2\right)\right]$.
The calculation of the length of one cycle can be done by the integral $L = a\displaystyle\int_0^{2\pi}\sqrt{1+\lambda^2 - 2\lambda\cos t}\,dt$. The shaded area in **Fig. 2.68** is $S = \pi a^2(2+\lambda^2)$.
The radius of curvature is $r = a\dfrac{(1+\lambda^2 - 2\lambda\cos t)^{3/2}}{\lambda(\cos t - \lambda)}$, which has the value $r_A = -a\dfrac{(1+\lambda)^2}{\lambda}$ at the maxima and the value $r_B = a\dfrac{(1-\lambda)^2}{\lambda}$ at the minima.

2.13.3 Epicycloid

A curve is called an *epicycloid*, if it is described by a point of the perimeter of a circle while this circle rolls along the outside of another circle without sliding **(Fig. 2.69)**. The equation of the epicycloid in parametric form is

$$x = (A+a)\cos\varphi - a\cos\frac{A+a}{a}\varphi, \quad y = (A+a)\sin\varphi - a\sin\frac{A+a}{a}\varphi \quad (-\infty < \varphi < \infty), \tag{2.244}$$

where A is the radius of the fixed circle, a is the radius of the rolling one, and φ is the angle $\sphericalangle C0x$. The shape of the curve depends on the quotient $m = \dfrac{A}{a}$.

For $m = 1$ one gets the *cardioid*.

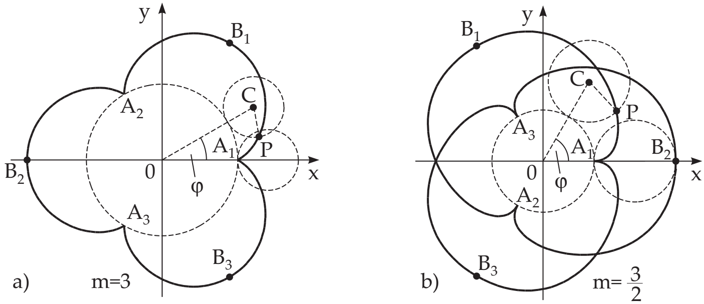


Figure 2.69

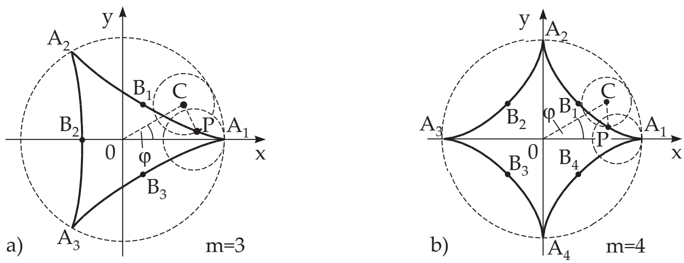


Figure 2.70

Case *m* integer: For an integer m the curve consists of m identically shaped branches surrounding the fixed curve **(Fig. 2.69a)**. The cusps $A_1, A_2, \ldots, A_m$ are at $\left(\rho = A,\ \ \varphi = \dfrac{2k\pi}{m}\ \ (k = 0, 1, \ldots, m-1)\right)$, the vertices $B_1, B_2, \ldots, B_m$ are at $\left(\rho = A + 2a,\ \ \ \varphi = \dfrac{2\pi}{m}\left(k + \dfrac{1}{2}\right)\right)$.

Case *m* a rational fraction: If m is a non-integer rational number, the identically shaped branches follow each other around the fixed circle, overlapping the previous ones, until the moving point P returns back to the starting-point after a finite number of circuits **(Fig. 2.69b)**.

Case *m* an irrational: For an irrational m the number of round trips is infinite; the point P never returns to the starting-point.

The length of one branch is $L_{A_1B_1A_2} = \dfrac{8(A+a)}{m}$. For an integer m the total length of the closed curve is $L_{\text{total}} = 8(A+a)$. The area of the sector $A_1B_1A_2A_1$ (without the sector of the fixed circle) is $S = \pi a^2\left(\dfrac{3A+2a}{A}\right)$. The radius of curvature is $r = \dfrac{4a(A+a)}{2a+A}\sin\dfrac{A\varphi}{2a}$, at the vertices $r_B = \dfrac{4a(A+a)}{2a+A}$.

2.13.4 Hypocycloid and Astroid

A curve is called a *hypocycloid* if it is described by a point of the perimeter of a circle, while this circle rolls along the inside of another circle without sliding **(Fig. 2.70)**. The equation of the hypocycloid,

the coordinates of the vertices, the cusps, the formulas for the arc-length, the area, and the radius of curvature are similar to the corresponding formulas for the epicycloid, only "$+a$" is to be replaced by "$-a$". The number of cusps for integer, rational or irrational m is the same as for the epicycloid (now $m > 1$ holds).

Case $m = 2$: For $m = 2$ the curve is actually the diameter of the fixed circle.

Case $m = 3$: For $m = 3$ the hypocycloid has three branches **(Fig. 2.70a)** with the equation

$$x = a(2\cos\varphi + \cos 2\varphi), \qquad y = a(2\sin\varphi - \sin 2\varphi) \tag{2.245a}$$

and $L_{\text{total}} = 16a$, $S_{\text{total}} = 2\pi a^2$.

Case $m = 4$: For $m = 4$ **(Fig. 2.70b)** the hypocycloid has four branches, and it is called an *astroid* (or *asteroid*). Its equation in Cartesian coordinates and in parametric form is

$$x^{2/3} + y^{2/3} = A^{2/3}, \quad (2.245\text{b}) \qquad x = A\cos^3\varphi, \quad y = A\sin^3\varphi \quad (0 \le \varphi < \pi) \quad (2.245\text{c})$$

and $L_{\text{total}} = 24a = 6A$, $S_{\text{total}} = \frac{3}{8}\pi A^2$.

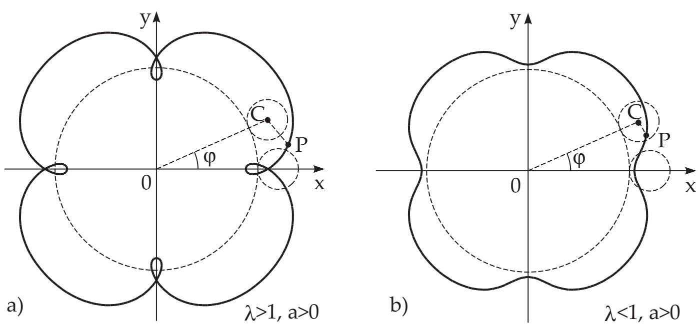


Figure 2.71

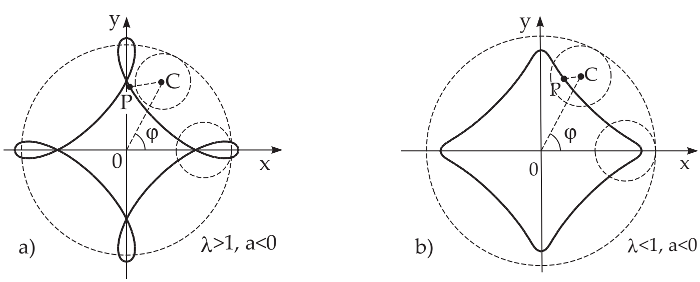


Figure 2.72

2.13.5 Prolate and Curtate Epicycloid and Hypocycloid

The *prolate* and *curtate epicycloid* and the *prolate* and *curtate hypocycloid*, which are also called the *epitrochoid* and *hypotrochoid*, are curves **(Fig. 2.71** and **Fig. 2.72)** described by a point, which is inside or outside of a circle, fixed on a half-line starting at the center of the circle, while the circle rolls around

the outside (epitrochoid) or the inside (hypotrochoid) of another circle, without sliding. The equation of the epitrochoid in parametric form is

$$x = (A + a)\cos\varphi - \lambda a \cos\left(\frac{A+a}{a}\varphi\right), \qquad y = (A + a)\sin\varphi - \lambda a \sin\left(\frac{A+a}{a}\varphi\right), \tag{2.246a}$$

where A is the radius of the fixed circle and a is the radius of the rolling one. For the hypocycloid "$+a$" is to be replaced by "$-a$". For $\lambda a = \overline{CP}$ one of the inequalities $\lambda > 1$ or $\lambda < 1$ is valid, depending on whether the prolate or the curtate curve is considered.

For $A = 2a$, and for arbitrary $\lambda \neq 1$ the hypocycloid with equation

$$x = a(1 + \lambda)\cos\varphi\,, \quad y = a(1 - \lambda)\sin\varphi \quad (0 \le \varphi < 2\pi) \tag{2.246b}$$

describes an ellipse with semi-axes $a(1 + \lambda)$ and $a(1 - \lambda)$. For $A = a$ it results in the Pascal limaçon (see also 2.12.3, p. 98):

$$x = a(2\cos\varphi - \lambda\cos 2\varphi), \quad y = a(2\sin\varphi - \lambda\sin 2\varphi). \tag{2.246c}$$

Remark: For the Pascal limaçon on 2.12.3, p. 98 the quantity denoted by a there is denoted by $2\lambda a$ here, and the l there is the diameter $2a$ here. Furthermore the coordinate system is different.

2.14 Spirals

2.14.1 Archimedean Spiral

An *Archimedean spiral* is a curve **(Fig. 2.73)** described by a point which is moving with constant speed v on a ray, while this ray rotates around the origin at a constant angular velocity ω. The equation of the Archimedean spiral in polar coordinates is

$$\rho = a|\varphi|, \quad a = \frac{v}{\omega} \quad (a > 0, -\infty < \varphi < \infty)\,. \tag{2.247}$$

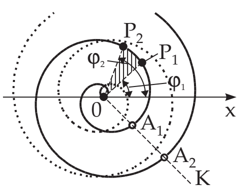


Figure 2.73

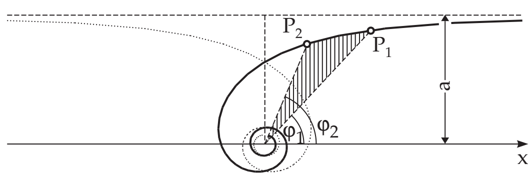


Figure 2.74

The curve has two branches ($\varphi < 0\,, \varphi > 0$) in a symmetric position with respect to the y-axis. Every ray $0K$ intersects the curve at the points 0, A_1, A_2,...,A_n,...; their distance is $\overline{A_iA_{i+1}} = 2\pi a$. The arclength $\overset{\frown}{0P}$ is $L = \frac{a}{2}\left(\varphi\sqrt{\varphi^2+1} + \operatorname{Arsinh}\varphi\right)$, where for large φ the expression $\frac{2L}{a\varphi^2}$ tends to 1. The area of the sector P_10P_2 is $S = \frac{a^2}{6}({\varphi_2}^3 - {\varphi_1}^3)$. The radius of curvature is $r = a\frac{(\varphi^2+1)^{3/2}}{\varphi^2+2}$ and at the origin $r_0 = \frac{a}{2}$.

2.14.2 Hyperbolic Spiral

The equation of the *hyperbolic spiral* in polar coordinates is

$$\rho = \frac{a}{|\varphi|} \quad (a > 0,\ -\infty < \varphi < 0,\ 0 < \varphi < \infty)\,. \tag{2.248}$$

The curve of the hyperbolic spiral **(Fig. 2.74)** has two branches ($\varphi < 0, \varphi > 0$) in a symmetric position with respect to the y-axis. The line $y = a$ is the asymptote for both branches, and the origin is an asymptotic point. The area of the sector $P_1 0 P_2$ is $S = \frac{a^2}{2}\left(\frac{1}{\varphi_1} - \frac{1}{\varphi_2}\right)$, and $\lim\limits_{\varphi_2 \to \infty} S = \frac{a^2}{2\varphi_1}$ is valid. The radius of curvature is $r = \frac{a}{\varphi}\left(\frac{\sqrt{1+\varphi^2}}{\varphi}\right)^3$.

2.14.3 Logarithmic Spiral

The *logarithmic spiral* is a curve **(Fig. 2.75)** which intersects all the rays starting at the origin 0 at the same angle α. The equation of the logarithmic spiral in polar coordinates is

$$\rho = a e^{k\varphi} \quad (a > 0,\ -\infty < \varphi < \infty), \tag{2.249}$$

where $k = \cot\alpha$. The origin is the asymptotic point of the curve. The arclength $\overset{\frown}{P_1P_2}$ is $L = \frac{\sqrt{1+k^2}}{k}(\rho_2 - \rho_1)$, the limit of the arclength $\overset{\frown}{0P}$ calculated from the origin is $L_0 = \frac{\sqrt{1+k^2}}{k}\rho$. The radius of curvature is $r = \sqrt{1+k^2}\rho = L_0 k$.

Special case of a circle: For $\alpha = \frac{\pi}{2}$ holds $k = 0$, and the curve is a circle.

2.14.4 Evolvent of the Circle

The *evolvent of the circle* is a curve **(Fig. 2.76)** which is described by the endpoint of a string while rolling it off a circle, and always keeping it tight, so that $\overset{\frown}{AB} = \overline{BP}$. The equation of the *evolvent of the circle* is in parametric form

$$x = a\cos\varphi + a\varphi\sin\varphi, \quad y = a\sin\varphi - a\varphi\cos\varphi, \tag{2.250}$$

where a is the radius of the circle, and $\varphi = \measuredangle B0x$. The curve has two branches in symmetric position with respect to the x-axis. It has a cusp at $A(a, 0)$, and the intersection points with the x-axis are at $x = \frac{a}{\cos\varphi_0}$, where φ_0 are the roots of the equation $\tan\varphi = \varphi$. The arclength of $\overset{\frown}{AP}$ is $L = \frac{1}{2}a\varphi^2$. The radius of curvature is $r = a\varphi = \sqrt{2aL}$; the centre of curvature B is on the circle, i.e., the circle is the evolute of the curve.

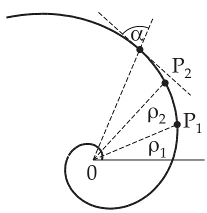


Figure 2.75

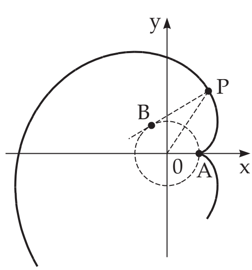


Figure 2.76

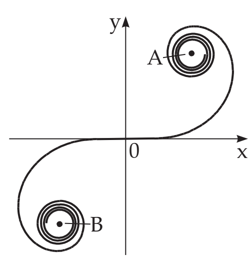


Figure 2.77

2.14.5 Clothoid

The *clothoid* is a curve **(Fig. 2.77)** such that at every point the radius of curvature is inversely proportional to the arclength between the origin and the considered point:

$$r = \frac{a^2}{s} \quad (a > 0)\,. \tag{2.251a}$$

The equation of the clothoid in parametric form is

$$x = a\sqrt{\pi}\int_0^t \cos\frac{\pi t^2}{2}\,dt, \quad y = a\sqrt{\pi}\int_0^t \sin\frac{\pi t^2}{2}\,dt \quad \text{with} \quad t = \frac{s}{a\sqrt{\pi}}, \quad s = \overset{\frown}{0P}\,. \tag{2.251b}$$

The integrals cannot be expressed in terms of elementary functions; but for any given value of the parameter $t = t_0, t_1, \dots$ it is possible to calculate them by numerical integration (see 19.3, p. 963), so the clothoid can be drawn pointwise. About calculations with a computer see the literature.

The curve is centrosymmetric with respect to the origin, which is also the inflection point. At the inflection point the x-axis is the tangent line. At A and B the curve has asymptotic points with coordinates $\left(+\frac{a\sqrt{\pi}}{2}, +\frac{a\sqrt{\pi}}{2}\right)$ and $\left(-\frac{a\sqrt{\pi}}{2}, -\frac{a\sqrt{\pi}}{2}\right)$. The clothoid is applied, for instance in road construction, where the transition between a line and a circular arc is made by a clothoid segment.

2.15 Various Other Curves

2.15.1 Catenary Curve

The *catenary curve* is a curve wich has the shape of a homogeneous, flexible but inextensible heavy chain hung at both ends **(Fig. 2.78)** represented by a continuous line. The equation of the *catenary curve* is

$$y = a\cosh\frac{x}{a} = a\frac{e^{x/a} + e^{-x/a}}{2} \quad (a > 0). \tag{2.252}$$

The parameter a determines the vertex A at $(0, a)$. The curve is symmetric to the y-axis, and is always higher than the parabola $y = a + \frac{x^2}{2a}$, which is represented by the broken line in **Fig. 2.78**. The arclength of $\overset{\frown}{AP}$ is $L = a\sinh\frac{x}{a} = a\frac{e^{x/a} - e^{-x/a}}{2}$. The area of the region $0APM$ has the value $S = a\,L = a^2\sinh\frac{x}{a}$. The radius of curvature is $r = \frac{y^2}{a} = a\cosh^2\frac{x}{a} = a + \frac{L^2}{a}$.

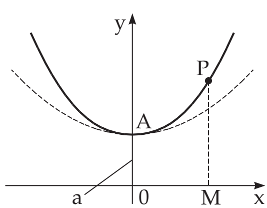


Figure 2.78

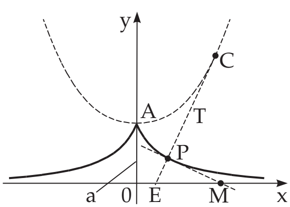


Figure 2.79

The catenary curve is the evolute of the tractrix (see 3.6.1.6, p. 254), so the tractrix is the evolvent (see 3.6.1.6, p. 255) of the catenary curve with vertex A at $(0, a)$.

2.15.2 Tractrix

The *tractrix* (the thick line in **Fig. 2.79**) is a curve such that the length of the segment $\overline{PM}$ of the tangent line between the point of contact P and the intersection point with a given straight line, here the x-axis, is a constant a. Fasting one end of an inextensible string of length a to a material point P, and dragging the other end along a straight line, here the x-axis, then P draws a tractrix. The equation of the tractrix is

$$x = a \operatorname{Arcosh} \frac{a}{y} \pm \sqrt{a^2 - y^2} = a \ln \frac{a \pm \sqrt{a^2 - y^2}}{y} \mp \sqrt{a^2 - y^2} \quad (a > 0\,, 0 < y \leq a). \tag{2.253}$$

The x-axis is the asymptote. The point A at $(0, a)$ is a cusp. The curve is symmetric with respect to the y-axis. The arclength of $\overset{\frown}{AP}$ is $L = a \ln \frac{a}{y}$. For increasing arclength L the difference $L - x$ tends to the value $a(1 - \ln 2) \approx 0.307a$, where x is the abscissa of the point P. The radius of curvature is $r = a \cot \frac{x}{y}$. The radius of curvature $\overline{PC}$ and the segment $\overline{PE} = b$ are inversely proportional: $rb = a^2$. The evolute (see 3.6.1.6, p. 254) of the tractrix, i.e., the geometric locus of the centers of circles of curvature C, is the catenary curve (2.252), represented by the dotted line in **Fig. 2.79**.

2.16 Determination of Empirical Curves

2.16.1 Procedure

2.16.1.1 Curve-Shape Comparison

If there are only empirical data for a function $y = f(x)$, it is possible to get an approximate formula in two steps. First choose a formula for an approximation which contains free parameters. Then calculate the values of the parameters. If there is no theoretical description for the type of formula, then first choose the approximate formula which is the simplest among the possible functions, comparing their curves with the curve of empirical data. Estimation of similarity by eye can be deceptive. Therefore, after the choice of an approximate formula, and before the determination of the parameters, it is to check whether it is appropriate.

2.16.1.2 Rectification

Supposing there is a definite relation between x and y and in the chosen approximate formula two functions $X = \varphi(x, y)$ and $Y = \psi(x, y)$ are introduced such that a linear relation of the form

$$Y = AX + B \tag{2.254}$$

holds, where A and B are constant. Calculating the corresponding X and Y values for the given x and y values, and considering their graphical representation, it is easy to check if they are approximately on a straight line, or not. Then it can be decided whether the chosen formula is appropriate.

■ **A:** If the approximate formula is $y = \frac{x}{ax + b}$, then substituting $X = x, Y = \frac{x}{y}$, one gets $Y = aX + b$. Another possible substitution is $X = \frac{1}{x}$, $Y = \frac{1}{y}$. Then $Y = a + bX$ follows.

■ **B:** Using semi-logarithmic paper, 2.17.2.1, p. 116.

■ **C:** Using double logarithmic paper, 2.17.2.2, p. 117.

In order to decide whether empirical data satisfy a linear relation $Y = AX + B$ or not, one can use *linear regression* or *correlation* (see 16.3.4, p. 839). The reduction of a functional relationship to a linear

relation is called *rectification*. Examples of rectification of some formulas are given in 2.16.2, p. 109, and for an example discussed in detail, see in 2.16, p. 114.

2.16.1.3 Determination of Parameters

The most important and most accurate method of determining the parameters is the *least squares method* (see 16.3.4.2, p. 841). In several cases, however, even simpler methods can be used with success, for instance the *mean value method*.

1. Mean Value Method

The *mean value method* uses the linear dependence of the "rectified" variables X and Y, i.e., $Y = AX + B$ as follows: The conditional equations $Y_i = AX_i + B$ for the given values Y_i, X_i are to be divided into two groups, which have the same size, or approximately the same size. By adding the equations in the groups one gets two equations, from which A and B can be determined. Then replacing X and Y by the original variables x and y again, one gets the connection between x and y, which is what one was looking for.
If not all the parameters are to be determined, one can apply the mean value method again with a rectification by other amounts $\overline{X}$ and $\overline{Y}$ (see, e.g., 2.16.2.11, p. 113).
Rectification and the mean value method are used above all when certain parameters occur in non-linear relations in an approximate formula, as for instance in (2.267b), (2.267c).

2. Least Squares Method

When certain parameters occur in non-linear relations in the approximation formula, the *least squares method* usually leads to a *non-linear fitting problem*. Their solution needs a lot of numerical calculations and also a good initial approximation. These approximations can be determined by the rectification and mean value method.

2.16.2 Useful Empirical Formulas

In this paragraph some of the simplest cases of empirical functional dependence are discussed, and the corresponding graphs are presented. Each figure shows several curves corresponding to different parameter values involved in the formula. The influence of the parameters upon the forms of the curves is discussed in the following sections.
For the choice of the appropriate function, usually only that part of the corresponding graph is to be considered, which is used for the reproduction of the empirical data. Therefore, e.g., one should not think that the formula $y = ax^2 + bx + c$ is suitable only in the case when the curve of the empirical data have a maximum or minimum.

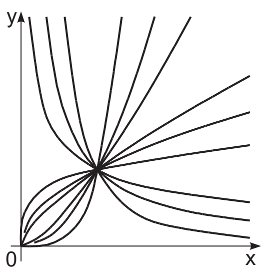


Figure 2.80

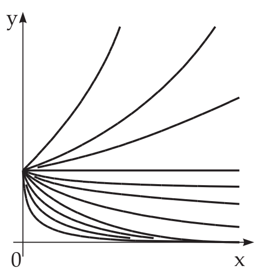


Figure 2.81

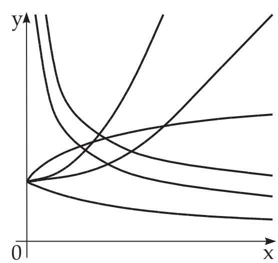


Figure 2.82

2.16.2.1 Power Functions

1. Type $y = ax^b$:

Typical shapes of curve for different values of the exponent b

$$y = ax^b \tag{2.255a}$$

are shown in **Fig. 2.80**. The curves for different values of the exponent are also represented in **Figs. 2.15, 2.21, 2.24, 2.25** and **Fig. 2.26**. The functions are discussed on pages 66, 70 and 71 for the formula (2.44) as a *parabola of order n*, formula (2.45) as a *reciprocal proportionality* and formula (2.50) as a *reciprocal power function*. The rectification is made by taking the logarithm

$$X = \log x, \quad Y = \log y: \quad Y = \log a + bX. \tag{2.255b}$$

2. Type $y = ax^b + c$:

The formula

$$y = ax^b + c \tag{2.256a}$$

produces the same curve as in (2.255a), but it is shifted by c in the direction of y **(Fig. 2.82)**. If b is given, the following rectification can be used:

$$X = x^b, \quad Y = y: \quad Y = aX + c. \tag{2.256b}$$

If b is not known first c is to be determined, then the rectification can be done in accordance with

$$X = \log x\,, \quad Y = \log(y - c): \quad Y = \log a + bx\,. \tag{2.256c}$$

In order to determine a first approach of c, one can choose two arbitrary abscissae x_1, x_2 and a third one, $x_3 = \sqrt{x_1 x_2}$ as well as the corresponding ordinates y_1, y_2, y_3, so that now

$$c = \frac{y_1 y_2 - {y_3}^2}{y_1 + y_2 - 2y_3} \tag{2.256d}$$

holds. After having determined a and b, one can get a corrected c, namely as the average of the amounts $y - ax^b$.

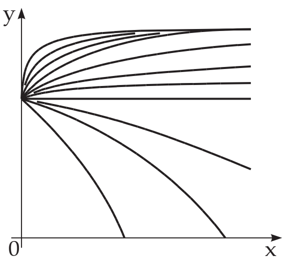


Figure 2.83

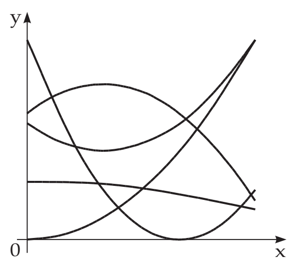


Figure 2.84

2.16.2.2 Exponential Functions

1. Type $y = ae^{bx}$:

The characteristic shapes of the curves of the function

$$y = ae^{bx} \tag{2.257a}$$

are shown in **Fig. 2.81**. The discussion of the *exponential function* (2.54) and its graph **(Fig. 2.26)** is presented in 2.6.1, p. 72. For the rectification one takes

$$X = x, \quad Y = \log y: \quad Y = \log a + b \log e \cdot X. \tag{2.257b}$$

2. Type $y = ae^{bx} + c$:

The formula

$$y = ae^{bx} + c \tag{2.258a}$$

produce the same curve as (2.257a), but it is shifted by c in the direction of y **(Fig. 2.83)**. Begin with the determination of a first approach c_1 for c then the rectification by logarithm:

$$Y = \log(y - c_1), \quad X = x: \quad Y = \log a + b \log e \cdot X. \tag{2.258b}$$

In order to determine c as for (2.256d) two arbitrary abscissae x_1, x_2 are to be chosen and $x_3 = \frac{x_1 + x_2}{2}$ as well as the corresponding ordinates y_1, y_2, y_3 to get $c = \frac{y_1 y_2 - {y_3}^2}{y_1 + y_2 - 2y_3}$. After having determined a and b one can get a corrected c, namely as the average of the amounts $y - ax^b$.

2.16.2.3 Quadratic Polynomial

Possible shapes of curves of the *quadratic polynomial*

$$y = ax^2 + bx + c \tag{2.259a}$$

are shown in **Fig. 2.84**. For the discussion of quadratic polynomials (2.41) and their curves **Fig. 2.12** (see 2.3.2, p. 64). Usually the coefficients a, b and c are determined by the least squares method; but in this case also rectification is possible. Choosing an arbitrary point of data (x_1, y_1) one rectifies

$$X = x, \quad Y = \frac{y - y_1}{x - x_1}: \quad Y = (b + ax_1) + aX. \tag{2.259b}$$

If the given x values form an arithmetical sequence with a difference h, one rectifies

$$Y = \Delta y, \quad X = x: \quad Y = (bh + ah^2) + 2ahX. \tag{2.259c}$$

In both cases after the determination of a and b from the equation

$$\sum y = a \sum x^2 + b \sum x + nc \tag{2.259d}$$

c is to be calculated; n is the number of the given x values, for which the sum is calculated.

2.16.2.4 Rational Linear Functions

The *rational linear function*

$$y = \frac{ax + b}{cx + d} \tag{2.260a}$$

is discussed in 2.4 with (2.46) and graphical representation **Fig. 2.17** (see p. 66).
Choosing an arbitrary data point (x_1, y_1) the rectification is done by the formulas

$$Y = \frac{x - x_1}{y - y_1}, \quad X = x: \quad Y = A + BX. \tag{2.260b}$$

After determining the values A and B the relation can be written in the form (2.260c). Sometimes instead of (2.260a) the form (2.260d)is sufficient:

$$y = y_1 + \frac{x - x_1}{A + Bx}, \ (2.260\text{c}) \qquad y = \frac{x}{cx + d} \quad \text{or} \quad y = \frac{1}{cx + d}. \tag{2.260d}$$

Then in the first case the rectification can be made by $X = \frac{1}{x}$ and $Y = \frac{1}{y}$ or $X = x$ and $Y = \frac{x}{y}$ and by $X = x$ and $Y = \frac{1}{y}$ in the second case.

2.16.2.5 Square Root of a Quadratic Polynomial

Several possible shapes of curves of the equation

$$y^2 = ax^2 + bx + c \tag{2.261}$$

are shown in **Fig. 2.85**. The discussion of the function (2.52) and its graph **Fig. 2.23** (see p. 71).
If introducing the new variable $Y = y^2$, the problem can be reduced to the case of the quadratic polynomial in 2.16.2.3, p. 111.

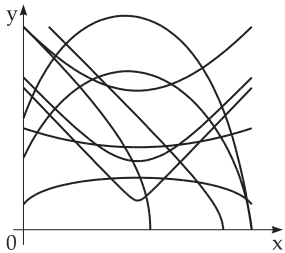


Figure 2.85

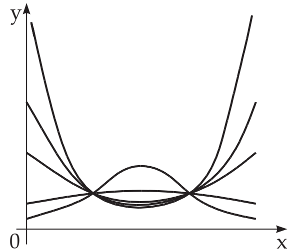


Figure 2.86

2.16.2.6 General Error Curve

The typical shapes of curves of the functions

$$y = ae^{bx+cx^2} \quad \text{or} \quad \log y = \log a + bx \log e + cx^2 \log e \tag{2.262}$$

are shown in **Fig. 2.86**. The discussion of the function with equation (2.61) and its graph **Fig. 2.31** (see p. 75).
Introducing the new variable $Y = \log y$, the problem is reduced to the case of the quadratic polynomial in 2.16.2.3, p. 111.

2.16.2.7 Curve of Order Three, Type II

The possible shapes of graphs of the function

$$y = \frac{1}{ax^2 + bx + c}. \tag{2.263}$$

are represented in **Fig. 2.87**. The discussion of the function with equation (2.48) and with graphs **Fig. 2.19** (see p. 67).

By introducing the new variable $Y = \frac{1}{y}$, the problem is reduced to the case of the quadratic polynomial in 2.16.2.3, p. 111.

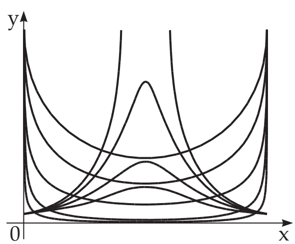


Figure 2.87

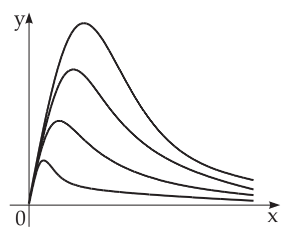


Figure 2.88

2.16.2.8 Curve of Order Three, Type III

Typical shapes of curves of functions of the type

$$y = \frac{x}{ax^2 + bx + c} \tag{2.264}$$

are represented in **Fig. 2.88**. The discussion of the function with equation (2.49) and with graphs **Fig. 2.20** (see p. 68).

Introducing the new variable $Y = \dfrac{x}{y}$ the problem can be reduced to the case of the quadratic polynomial in 2.16.2.3, p. 111.

2.16.2.9 Curve of Order Three, Type I

Typical shapes of curves of functions of the type

$$y = a + \frac{b}{x} + \frac{c}{x^2} \tag{2.265}$$

are represented in **Fig. 2.89**. The discussion of the function with equation (2.47) and with graphs **Fig. 2.18** (see p. 67).

Introducing the new variable $X = \dfrac{1}{x}$ the problem can be reduced to the case of the quadratic polynomial in 2.16.2.3, p. 111.

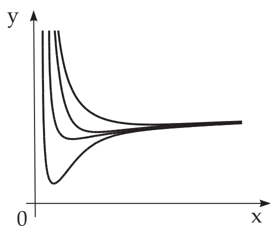


Figure 2.89

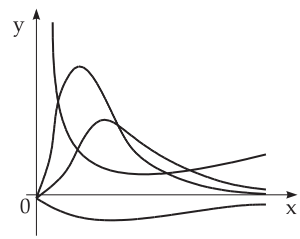


Figure 2.90

2.16.2.10 Product of Power and Exponential Functions

Typical shapes of curves of functions of the type

$$y = ax^b e^{cx} \tag{2.266a}$$

are represented in **Fig. 2.90**. The discussion of the function with equation (2.62) and with graphs **Fig. 2.31** (see p. 75).
If the empirical values of x form an arithmetical sequence with difference h, the rectification follows in accordance with

$$Y = \Delta \log y, \quad X = \Delta \log x: \quad Y = hc \log e + bX. \tag{2.266b}$$

Here $\Delta \log y$ and $\Delta \log x$ denote the difference of two subsequent values of $\log y$ and $\log x$ respectively.
If the x values form a geometric sequence with quotient q, then the rectification follows by

$$X = x, \quad Y = \Delta \log y: \quad Y = b \log q + c(q-1)X \log e. \tag{2.266c}$$

After b and c are determined the logarithm of the given equation is taken, and the value of $\log a$ is calculated like in (2.259d).
If the given x values do not form a geometric sequence, but one can choose pairs of two values of x such that their quotient q is the same constant, then the rectification is the same as in the case of a geometric sequence of x values with the substitution $Y = \Delta_1 \log y$. Here $\Delta_1 \log y$ denotes the difference of the two values of $\log y$ whose corresponding x values result in the constant quotient q (see 2.16.2.12, p. 114).

2.16.2.11 Exponential Sum

Typical shapes of curves of the *exponential sum*

$$y = ae^{bx} + ce^{dx} \tag{2.267a}$$

are represented in **Fig. 2.91**. The discussion of the function with equation (2.60) and with graphs **Fig. 2.29** (see p. 74).
If the values of x form an arithmetical sequence with difference h, and y, y_1, y_2 are any three consecutive values of the given function, then the rectification is made by

$$Y = \frac{y_2}{y}, \quad X = \frac{y_1}{y}: \quad Y = (e^{bh} + e^{dh})X - e^{bh} \cdot e^{dh}. \tag{2.267b}$$

After b and d are determined, follows again a rectification by

$$\overline{Y} = ye^{-dx}, \quad \overline{X} = e^{(b-d)x}: \quad \overline{Y} = a\overline{X} + c. \tag{2.267c}$$

2.16.2.12 Numerical Example

Find an empirical formula to describe the relation between x and y for given values in **Table 2.9**.

Choice of the Approximation Function: Comparing the graph prepared from the given data (**Fig. 2.92**) with the curves discussed before, one can see that formulas (2.264) or (2.266a) with curves in **Fig. 2.88** and **Fig. 2.90** can fit the considered case.

Determination of Parameters: Using the formula (2.264) to rectify are $\Delta\frac{x}{y}$ and x. The calculation shows, however, the relationship between x and $\Delta\frac{x}{y}$ is far from linear. To verify whether the formula (2.266a) is suitable one plots the graph of the relation between $\Delta \log x$ and $\Delta \log y$ for $h = 0,1$ in **Fig. 2.93**, and also between $\Delta_1 \log y$ and x for $q = 2$ in **Fig. 2.94**. In both cases the points fit a straight line well enough, so the formula $y = ax^b e^{cx}$ can be used.

Table 2.9 For the approximate determination of an empirically given function relation

x	y	$\frac{x}{y}$	$\Delta\frac{x}{y}$	$\lg x$	$\lg y$	$\Delta \lg x$	$\Delta \lg y$	$\Delta_1 \lg y$	y_{err}
0.1	1.78	0.056	0.007	−1.000	0.250	0.301	0.252	0.252	1.78
0.2	3.18	0.063	0.031	−0.699	0.502	0.176	+0.002	−0.097	3.15
0.3	3.19	0.094	0.063	−0.523	0.504	0.125	−0.099	−0,447	3.16
0.4	2.54	0.157	0.125	−0.398	0.405	0.097	−0.157	−0.803	2.52
0.5	1.77	0.282	0.244	−0.301	0.248	0.079	−0.191	−1.134	1.76
0.6	1.14	0.526	0.488	−0.222	0.057	0.067	−0.218	−1.455	1.14
0.7	0.69	1.014	0.986	−0.155	−0.161	0.058	−0.237	–	0.70
0.8	0.40	2.000	1.913	−0.097	−0.398	0.051	−0.240	–	0.41
0.9	0.23	3.913	3.78	−0.046	−0.638	0.046	−0.248	–	0.23
1.0	0.13	7.69	8.02	0.000	−0.886	0.041	−0.269	–	0.13
1.1	0.07	15.71	14.29	0.041	−1.155	0.038	−0.243	–	0.07
1.2	0.04	30.0	–	0.079	−1.398	–	–	–	0.04

In order to determine the constants a, b and c, a linear relation between x and $\Delta_1 \log y$ is to be searched by the method of mean values. Adding the conditional equations $\Delta_1 \log y = b \log 2 + cx \log e$ in groups of three equations each, yields

$$-0.292 = 0.903b + 0.2606c\,, \quad -3.392 = 0.903b + 0.6514c,$$

and from here $b = 1.966$ and $c = -7.932$ holds. To determine a, the equations of the form $\log y = \log a + b \log x + c \log e \cdot x$ are to be added, which yields $-2.670 = 12 \log a - 6.529 - 26.87$, so from $\log a = 2.561$, $a = 364$ follows. The values of y calculated from the formula $y = 364x^{1{,}966}e^{-7.032x}$ are given in the last column of **Table 2.9**; they are denoted by y_{err} as an approximation of y. The error sum of squares is 0.0024.
Using the parameters determined by rectification as initial values for the iterative solution of the non-

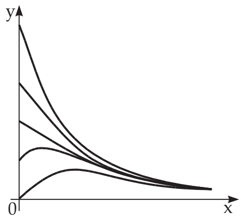


Figure 2.91

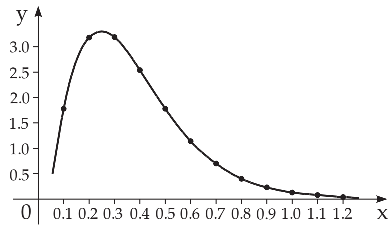


Figure 2.92

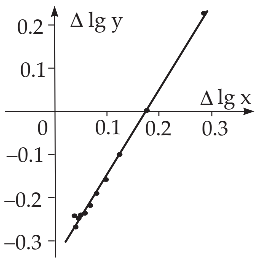


Figure 2.93

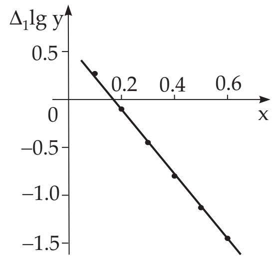


Figure 2.94

linear least squares problem (see 19.6.2.4, p. 987)

$$\sum_{i=1}^{12}[y_i - ax_i^b e^{cx_i}]^2 = \text{min!}$$

yields $a = 396.601\,986$, $b = 1.998\,098$, $c = -8.000\,0916$ with the small error sum of squares 0.000 0916.

2.17 Scales and Graph Paper

2.17.1 Scales

The base of a scale is a function $y = f(x)$. The task is to construct a *scale* from this function so that on a curve, e.g. on a line, the function values of y are to be inserted as the values of the argument x. A scale can be considered as a one-dimensional representation of a table of values.
The *scale equation* for the function $y = f(x)$ is:

$$y = l[f(x) - f(x_0)]. \tag{2.268}$$

The starting point of the scale is fixed at x_0. The *scale factor* l takes into consideration that for a concrete scale there it is only one given scale length.

■ **A Logarithmic Scale:** For $l = 10$ cm and $x_0 = 1$ the scale equation is
$y = 10[\lg x - \lg 1] = 10 \lg x$ (in cm). For the table of values

x	1	2	3	4	5	6	7	8	9	10
$y = \lg x$	0	0.30	0.48	0.60	0.70	0.78	0.85	0.90	0.95	1

one gets the scale shown in **Fig. 2.95**.

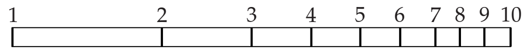


Figure 2.95

■ **B Slide Rule:** The most important application of the logarithmic scale, from a historical viewpoint, was the *slide rule*. Here, for instance, multiplication and division were performed with the help of two identically calibrated logarithmic scales, which can be shifted along each other.
From **Fig. 2.96** one can read: $y_3 = y_1 + y_2$, i.e., $\lg x_3 = \lg x_1 + \lg x_2 = \lg x_1 x_2$, hence $x_3 = x_1 \cdot x_2$; $y_1 = y_3 - y_2$, i.e., $\lg x_1 = \lg x_3 - \lg x_2 = \lg \frac{x_3}{x_2}$, so $x_1 = \frac{x_3}{x_2}$.

■ **C Volume Scale** on the lateral surface of a conical shaped funnel: A scale is to be marked on the funnel, so that the volume inside could be read from it. The data of the funnel are: Height $H = 15$ cm, upper diameter $D = 10$ cm.

Taking in mind **Fig. 2.97a** gives the scale equation as follows: Volume $V = \frac{1}{3}r^2\pi h$, apothem $s = \sqrt{h^2 + r^2}$, $\tan\alpha = \frac{r}{h} = \frac{D/2}{H} = \frac{1}{3}$. From these $h = 3r$, $s = r\sqrt{10}$, $V = \frac{\pi}{(\sqrt{10})^3}$ follows, so the scale equation is $s = \frac{\sqrt{10}}{\sqrt[3]{\pi}}\sqrt[3]{V} \approx 2.16\sqrt[3]{V}$. The following table of values contains the calibration marks on the funnel as in the figure:

V	0	50	100	150	200	250	300	350
s	0	7.96	10.03	11.48	12.63	13.61	14.46	15.22

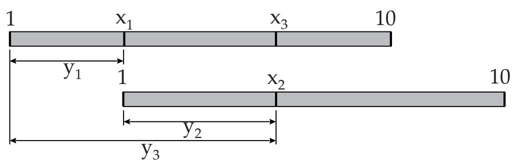


Figure 2.96

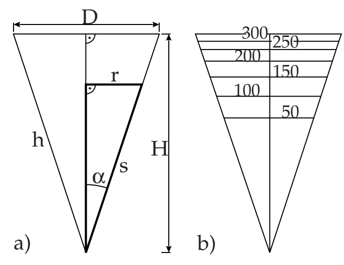


Figure 2.97

2.17.2 Graph Paper

The most useful graph paper is prepared so that the axes of a right-angle coordinate system are calibrated by the scale equations

$$x = l_1[g(u) - g(u_0)], \quad y = l_2[f(v) - f(v_0)]. \tag{2.269}$$

Here l_1 and l_2 are the scale factors; u_0 and v_0 are the initial points of the scale.

2.17.2.1 Semilogarithmic Paper

If the x-axis has an equidistant subdivision, and the y-axis has a logarithmic one, then one talks about *semilogarithmic paper* or about a *semilogarithmic coordinate system*.

Scale Equations:

$$x = l_1[u - u_0] \quad \text{(linear scale)}, \qquad y = l_2[\lg v - \lg v_0] \quad \text{(logarithmic scale)}. \tag{2.270}$$

The **Fig. 2.98** shows an example of semilogarithmic paper.

Representation of Exponential Functions: On semilogarithmic paper the graph of the exponential function

$$y = \alpha e^{\beta x} \quad (\alpha, \beta \text{ const}) \tag{2.271a}$$

is a straight line (see rectification in 2.16.2.2, p. 110). This property can be used in the following way: If the measuring points, introduced on semilogarithmic paper, lie approximately on a line, it can be supposed a relation between the variables as in (2.271a). With this line, estimated by eye, one can determine the approximate values of α and β: Considering two points $P_1(x_1, y_1)$ and $P_2(x_2, y_2)$ from this line one gets

$$\beta = \frac{\ln y_2 - \ln y_1}{x_2 - x_1} \quad \text{and, e.g.,} \quad \alpha = y_1 e^{\beta x_1}. \tag{2.271b}$$

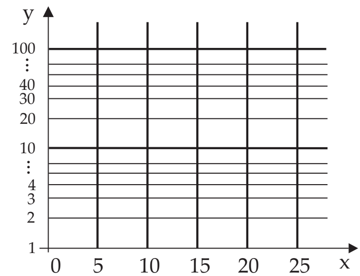


Figure 2.98

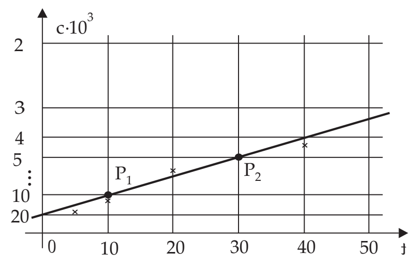


Figure 2.99

2.17.2.2 Double Logarithmic Paper

If both axes of a right-angle x, y coordinate system are calibrated with respect to the logarithm function, then one talks about *double logarithmic paper* or *log–log paper* or a *double logarithmic coordinate system*.

Scale Equations: The scale equations are

$$x = l_1[\lg u - \lg u_0], \qquad y = l_2[\lg v - \lg v_0], \tag{2.272}$$

where l_1, l_2 are the scale factors and u_0, v_0 are the initial points.

Representation of Power Functions (see 2.5.3, p. 71): Log–log paper has a similar arrangement to semilogarithmic paper, but the x-axis also has a logarithmic subdivision. In this coordinate system the graph of the power function

$$y = \alpha x^{\beta} \quad (\alpha, \beta \text{ const}) \tag{2.273}$$

is a straight line (see rectification of a power function in 2.16.2.1, p. 109). This property can be used in the same way as in the case of semilogarithmic paper.

2.17.2.3 Graph Paper with a Reciprocal Scale

The subdivisions of the scales on the coordinate axes follow from (2.45) for the function of inverse proportionality (see 2.4.1, p. 66).

Scale Equations: $x = l_1[u - u_0], \quad y = l_2\left[\frac{a}{v} - \frac{a}{v_0}\right] \quad (a \text{ const}),$ (2.274)

where l_1 and l_2 are the scale factors, and u_0, v_0 are the starting points.

■ **Concentration in a Chemical Reaction:** For a chemical reaction, the concentration denoted by $c = c(t)$, has been measured as a function of the time t, giving the following results for c:

t/min	5	10	20	40
$c\cdot 10^3/\text{mol/l}$	15.53	11.26	7.27	4.25

It can be supposed that the reaction is of second order, i.e., the relation should be

$$c(t) = \frac{c_0}{1 + c_0 kt} \quad (c_0, k \text{ const}). \tag{2.275}$$

Taking the reciprocal value of both sides, one gets $\frac{1}{c} = \frac{1}{c_0} + kt$, i.e., (2.275) can be represented as a line, if the corresponding graph paper has a reciprocal subdivision on the y-axis and a linear one on the x-axis. The scale equation for the y-axis is, e.g., $y = 10 \cdot \frac{1}{v}$ cm.

It is obvious from the corresponding **Fig. 2.99** that the measuring points lie approximately on a line, i.e., the supposed relation (2.275) is acceptable.
From these points the approximate values of both parameters k (reaction rate) and c_0 (initial concentration) can be determined. Choosing two points, e.g., $P_1(10, 10)$ and $P_2(30, 5)$, one gets:

$$k = \frac{1/c_1 - 1/c_2}{t_2 - t_1} \approx 0.005, \quad c_0 \approx 20 \cdot 10^{-3}.$$

2.17.2.4 Remark

There are several other possibilities for constructing and using graph papers. Although today in most cases there are high-capacity computers to analyze empirical data and measurement results, in everyday laboratory practice, when getting only a few data, graph papers are used quite often to show the functional relations and approximate parameter values needed as initial data for applied numerical methods (see the non-linear least squares method in 19.6.2.4, p. 987).

2.18 Functions of Several Variables

2.18.1 Definition and Representation

2.18.1.1 Representation of Functions of Several Variables

A variable value u is called a function of n *independent variables* $x_1, x_2, \ldots, x_n$, if for given values of the independent variables, u is a uniquely defined value. Depending on how many variables there are, two, three, or n, one writes

$$u = f(x, y), \quad u = f(x, y, z), \quad u = f(x_1, x_2, \ldots, x_n). \tag{2.276}$$

Substituting given numbers for the n independent variables yields a value system of the variables, which can be considered as a *point of the n-dimensional space*. The single independent variables are called arguments; sometimes the entire n *tuple* together is called the argument of the function.

Examples of Values of Functions:

■ **A:** $u = f(x, y) = xy^2$ has for the value system $x = 2$, $y = 3$ the value $f(2, 3) = 2 \cdot 3^2 = 18$.
■ **B:** $u = f(x, y, z, t) = x \ln(y - zt)$ takes for the value system $x = 3$, $y = 4$, $z = 3$, $t = 1$ the value $f(3, 4, 3, 1) = 3 \ln(4 - 3 \cdot 1) = 0$.

2.18.1.2 Geometric Representation of Functions of Several Variables

1. Representation of the Value System of the Variables

The value system of an argument of two variables x and y can be represented as a point of the plane given by Cartesian coordinates x and y. A value system of three variables x, y, z corresponds to a point given by the coordinates x, y, z in a three-dimensional Cartesian coordinate system. Systems of four or more coordinates cannot be represented obviously in our three-dimensional imagination.

Similarly to the three-dimensional case the system of n variables $x_1, x_2, \ldots, x_n$ is to be considered as a

point of the n-dimensional space given by Cartesian coordinates x_1, x_2,. . . ,x_n. In the above example **B**, the four variables define a point in four-dimensional space, with coordinates $x = 3$, $y = 4$, $z = 3$ and $t = 1$.

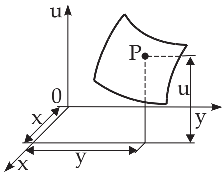


Figure 2.100

2. Representation of the Function $u = f(x, y)$ of Two Variables

a) A function of two independent variables can be represented by a surface in three-dimensional space, similarly to the graph representation of functions of one variable (**Fig. 2.100**, see also 3.6.3, p. 261). Considering the values of the independent variables of the domain as the first two coordinates, and the value of the function $u = f(x, y)$ as the third coordinate of a point in a Cartesian coordinate system, these points form a surface in three-dimensional space.

Examples of Surfaces of Functions:

■ **A:** $u = 1 - \frac{x}{2} - \frac{y}{3}$ represents a plane (**Fig. 2.101a**, see also 3.5.3.10, p. 218).

■ **B:** $u = \frac{x^2}{2} + \frac{y^2}{4}$ represents an elliptic paraboloid (**Fig. 2.101b**, see also 3.5.3.13, **5.**, p. 226).

■ **C:** $u = \sqrt{16 - x^2 - y^2}$ represents a hemisphere with $r = 4$ **(Fig. 2.101c)**.

b) The shape of the surface of the function $u = f(x, y)$ can be pictured with the help of intersection curves, which can be get by intersecting the surface parallel to the coordinate planes. The intersection curves $u =$ const are called *level curves*.

■ In **Fig. 2.101b,c** the level curves are ellipses or concentric circles (not denoted in the figure).

Remark: A function with an argument of three or more variables cannot be represented in three-dimensional space. Similarly to surfaces in three-dimensional space also the notion of a *hyper surface* in n-dimensional space is in use.

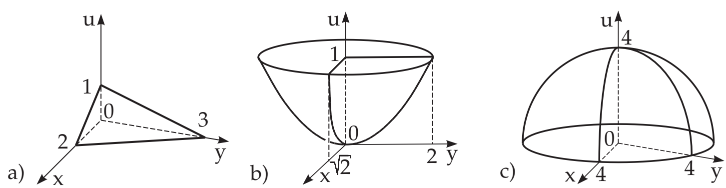


Figure 2.101

2.18.2 Different Domains in the Plane

2.18.2.1 Domain of a Function

The *domain of definition of a function* (or domain of a function) is the *set of the system of values* or *points* which can be taken by the variables of the argument of the function. The domains defined in this way can be very different. Mostly they are bounded or unbounded connected *sets of points*. Depending on whether the boundary belongs to the domain or not, the domain is closed or open. An open, connected set of points is called a *domain*. If the boundary belongs to the domain, it is called a *closed domain*, if it does not, sometimes it is called an *open domain*.

2.18.2.2 Two-Dimensional Domains

Fig. 2.102 shows the simplest cases of connected sets of points of two variables and their notation. Domains are represented here as the shaded part; closed domains, i.e., domains whose boundary belongs to

them, are bounded by thick curves in the figures; open domains are bounded by dotted curves. Including the entire plane there are only *simply connected domains* or *simply connected regions* in **Fig. 2.102**.

2.18.2.3 Three or Multidimensional Domains

These are handled similarly to the two-dimensional case. It concerns also the distinction between simply and multiply connected domains. Functions of more than three variables will be geometrically represented in the corresponding n-dimensional space.

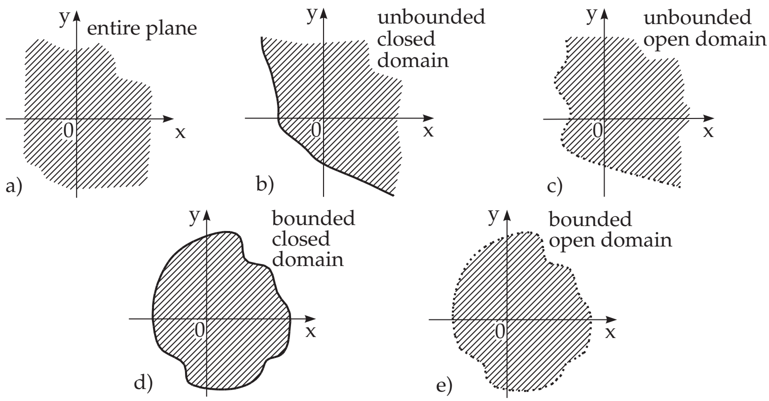


Figure 2.102

2.18.2.4 Methods to Determine a Function

1. Definition by Table of Values Functions of several variables can be defined by a table of values. An example of functions of two independent variables are the tables of values of elliptic integrals (see 21.9, p. 1103). The values of the independent variables are denoted on the top and on the left-hand side of the table. The required value of the function is in the intersection of the corresponding row and column. It is called a *table with double entry*.

2. Definition by Formulas Functions of several variables can be defined by one or more formulas.

■ **A:** $u = xy^2$.

■ **B:** $u = x\ \ln(y - zt)$.

■ **C:** $u = \begin{cases} x + y \ \mathit{for}\ x \geq 0,\ y \geq 0, \\ x - y \ \mathit{for}\ x \geq 0,\ y < 0, \\ -x + y \ \mathit{for}\ x < 0,\ y \geq 0, \\ -x - y \ \mathit{for}\ x < 0,\ y < 0. \end{cases}$

3. Domain of a Function Given by One Formula In the analysis mostly such functions are considered which are defined by formulas. Here the union of all value systems for which the analytical expression has a meaning is considered to be the domain, i.e., for which the expression has a unique, finite, real value.

Examples for Domains:

■ **A:** $u = x^2 + y^2$: The domain is the entire plane.

■ **B:** $u = \dfrac{1}{\sqrt{16 - x^2 - y^2}}$: The domain consists of all value systems x, y, satisfying the inequality $x^2 + y^2 < 16$. Geometrically this domain is the interior of the circle in **Fig. 2.103a**, an open domain.

■ **C:** $u = \arcsin(x + y)$: The domain consists of all value systems x, y, satisfying the inequality $-1 \leq$

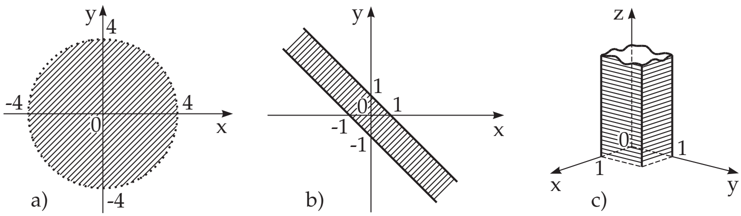


Figure 2.103

$x + y \leq +1$, i.e., the domain of the function is a closed domain, the stripe between the two parallel lines in **Fig. 2.103b**.

■ **D:** $u = \arcsin(2x - 1) + \sqrt{1 - y^2} + \sqrt{y} + \ln z$: The domain consists of the value system x, y, z, satisfying the inequalities $0 \leq x \leq 1$, $0 \leq y \leq 1$, $z > 0$, i.e., it consists of all points of the three dimensional x, y, z space lying above a square with side-length 1 shown in **Fig. 2.103c**.

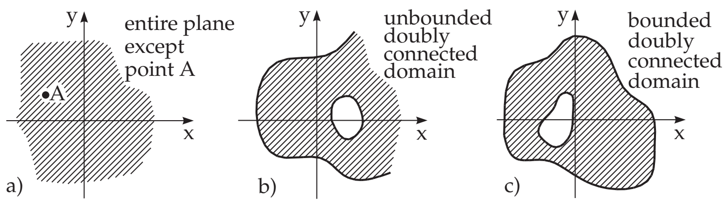


Figure 2.104

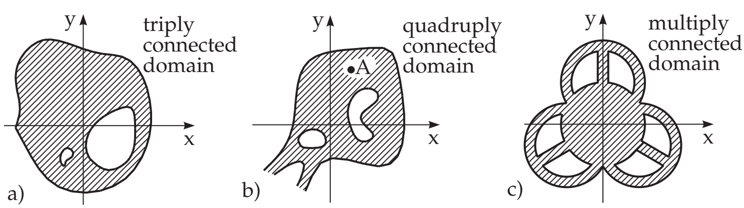


Figure 2.105

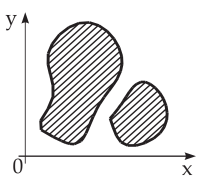


Figure 2.106

If from the interior of the considered part of the plane a point or a bounded, simply connected point set is missing, as shown in **Fig. 2.104**, it is called a *doubly-connected domain* or *doubly-connected region*. *Multiply connected domains* are represented in **Fig. 2.105**. A *non-connected region* is shown in **Fig. 2.106**.

2.18.2.5 Various Forms for the Analytical Representation of a Function

Functions of several variables can be defined in different ways, just as functions of one variable.

1. Explicit Representation
A function is given or defined in an explicit way if its value (the dependent variable) can be expressed by the independent variables:

$$u = f(x_1, x_2, \ldots, x_n)\,. \tag{2.277}$$

2. Implicit Representation
A function is given or defined in an implicit way if the relation between its value and the independent variables is given in the form:

$$F(x_1, x_2, \ldots, x_n, u) = 0\,, \tag{2.278}$$

if there is a unique value of u satisfying this equality.

3. Parametric Representation
A function is given in parametric form if the n arguments and the function are defined by n new variables, the parameters, in an explicit way, supposing there is a one-to-one correspondence between the parameters and the arguments. For a two-variable function, for instance

$$x = \varphi(r, s)\,, \quad y = \psi(r, s)\,, \quad u = \chi(r, s)\,, \tag{2.279a}$$

and for a three-variable function

$$x = \varphi(r, s, t)\,, \quad y = \psi(r, s, t)\,, \quad z = \chi(r, s, t)\,, \quad u = \kappa(r, s, t) \tag{2.279b}$$

etc.

4. Homogeneous Functions
A function $f(x_1, x_2, \ldots, x_n)$ of several variables is called a homogeneous function if the relation

$$f(\lambda x_1, \lambda x_2, \ldots, \lambda x_n) = \lambda^m f(x_1, x_2, \ldots, x_n) \tag{2.280}$$

holds for arbitrary λ. The number m is the *degree of homogeneity.*

■ **A:** For $u(x, y) = x^2 - 3xy + y^2 + x\sqrt{xy + \frac{x^3}{y}}$, the degree of homogeneity is $m = 2$.

■ **B:** For $u(x, y) = \frac{x + z}{2x - 3y}$, the degree of homogeneity is $m = 0$.

2.18.2.6 Dependence of Functions

1. Special Case of Two Functions
Two functions of two variables $u = f(x, y)$ and $v = \varphi(x, y)$, with the same domain G, are called *dependent functions* if one of them can be expressed as a function of the other one $u = F(v)$. For every point of the domain G of the functions the identity

$$f(x, y) = F(\varphi(x, y)) \quad \text{or} \quad \Phi(f, \varphi) = 0 \tag{2.281}$$

holds. If there is no such function $F(\varphi)$ or $\Phi(f, \varphi)$, one speaks about *independent functions.*

■ $u(x, y) = (x^2 + y^2)^2$, $v = \sqrt{x^2 + y^2}$ are defined in the domain $x^2 + y^2 \geq 0$, i.e., in the whole x, y plain, and they are dependent, because $u = v^4$ holds.

2. General Case of Several Functions
Similarly to the case of two functions, the m functions $u_1, u_2, \ldots, u_m$ of n variables $x_1, x_2, \ldots, x_n$ in their common domain G are called dependent if one of them can be expressed as a function of the others, i.e., if for every point of the domain G the identity

$$u_i = f(u_1, u_2, \ldots, u_{i-1}, u_{i+1}, \ldots, u_m) \quad \text{or} \quad \Phi(u_1, u_2, \ldots, u_m) = 0 \tag{2.282}$$

is valid. If there is no such functional relationship, they are independent functions.

■ The functions $u = x_1 + x_2 + \cdots + x_n$, $v = {x_1}^2 + {x_2}^2 + \cdots + {x_n}^2$ and $w = x_1x_2 + x_1x_3 + \cdots + x_1x_n + x_2x_3 + \cdots + x_{n-1}x_n$ are dependent because $v = u^2 - 2w$ holds.

3. Analytical Conditions for Independence

Suppose every partial derivative mentioned below exists. Two functions $u = f(x, y)$ and $v = \varphi(x, y)$ are independent on a domain if their *functional determinant* or *Jacobian determinant* is not identically zero here. Analogously, in the case of n functions of n variables $u_1 = f_1(x_1, \ldots, x_n), \ldots, u_n = f_n(x_1, \ldots, x_n)$:

$$\begin{vmatrix} \frac{\partial f}{\partial x} & \frac{\partial f}{\partial y} \\ \frac{\partial \varphi}{\partial x} & \frac{\partial \varphi}{\partial y} \end{vmatrix}, \text{ short } \frac{D(f,\varphi)}{D(x,y)} \text{ or } \frac{D(u,v)}{D(x,y)}, \tag{2.283a}$$

$$\begin{vmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \ldots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \ldots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial f_n}{\partial x_1} & \frac{\partial f_n}{\partial x_2} & \ldots & \frac{\partial f_n}{\partial x_n} \end{vmatrix} \equiv \frac{D(f_1, f_2, \ldots, f_n)}{D(x_1, x_2, \ldots, x_n)} \neq 0\,. \tag{2.283b}$$

If the number m of the functions $u_1, u_2, \ldots, u_m$ is smaller than the number of variables $x_1, x_2, \ldots, x_n$, these functions are independent if at least one subdeterminant of order m of the matrix (2.283c) is not identically zero:

$$\begin{pmatrix} \frac{\partial u_1}{\partial x_1} & \frac{\partial u_1}{\partial x_2} & \ldots & \frac{\partial u_1}{\partial x_n} \\ \frac{\partial u_2}{\partial x_1} & \frac{\partial u_2}{\partial x_2} & \ldots & \frac{\partial u_2}{\partial x_n} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial u_m}{\partial x_1} & \frac{\partial u_m}{\partial x_2} & \ldots & \frac{\partial u_m}{\partial x_n} \end{pmatrix}. \tag{2.283c}$$

The number of independent functions is equal to the rank r of the matrix (2.283c) (see 4.1.4, **7.**, p. 274). Here these functions are independent, whose derivatives are the elements of the non-vanishing determinant of order r. If $m > n$ holds, then among the given m functions at most n can be independent.

2.18.3 Limits

2.18.3.1 Definition

A function of two variables $u = f(x, y)$ has a limit A at $x = a$, $y = b$ if when x and y are arbitrarily close to a and b, respectively, then the value of the function $f(x, y)$ approaches arbitrarily closely the value A. Then one writes:

$$\lim_{\substack{x \to a \\ y \to b}} f(x, y) = A\,. \tag{2.284}$$

The function may not be defined at (a, b), or if it is defined here, may not have the value A.

2.18.3.2 Exact Definition

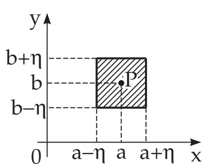


Figure 2.107

A function of two variables $u = f(x, y)$ has at the point (a, b) a limit

$$A = \lim_{\substack{x \to a \\ y \to b}} f(x, y) \tag{2.285a}$$

if for arbitrary positive ε there is a positive η such that (see **Fig. 2.107**)

$$|f(x, y) - A| < \varepsilon \tag{2.285b}$$

holds for every point (x, y) of the square

$$|x - a| < \eta, \quad |y - b| < \eta\,. \tag{2.285c}$$

2.18.3.3 Generalization for Several Variables

a) The notion of limit of a function of several variables can be defined analogously to the case of two variables.

b) The criteria for the existence of a limit of a function of several variables can be obtained by generalization of the criterion for functions of one variable, i.e., by reducing to the limit of a sequence or by applying the Cauchy condition for convergence (see 2.1.4.3, p. 53).

2.18.3.4 Iterated Limit

If for a function of two variables $f(x, y)$ first the limit for $x \to a$ has been determined, i.e., $\lim\limits_{x\to a} f(x, y)$ for constant y, then for the function obtained, which is now a function only of y, one determines the limit for $y \to b$, then the resulting number

$$B = \lim_{y\to b}\left(\lim_{x\to a} f(x, y)\right) \tag{2.286a}$$

is called an *iterated limit*. Changing the order of calculations generally yields an other limit:

$$C = \lim_{x\to a}\left(\lim_{y\to b} f(x, y)\right). \tag{2.286b}$$

In general $B \neq C$ holds, even if both limits exist.

■ For the function $f(x, y) = \dfrac{x^2 - y^2 + x^3 + y^3}{x^2 + y^2}$ for $x \to 0$, $y \to 0$ one gets the iterated limits $B = -1$ and $C = +1$.

Remark: If the function $f(x, y)$ has a limit $A = \lim\limits_{\substack{x\to a\\ y\to b}} f(x, y)$, and both B and C exist, then $B = C = A$ is valid. The existence of B and C does not follow from the existence of A. From the equality of the limits $B = C$ the existence of the limit A does not follow.

2.18.4 Continuity

A function of two variables $f(x, y)$ is continuous at $x = a$, $y = b$, i.e., at the point (a, b), if 1. the point (a, b) belongs to the domain of the function and 2. the limit for $x \to a$, $y \to b$ exists and is equal to the value, i.e.,

$$\lim_{\substack{x\to a\\ y\to b}} f(x, y) = f(a, b). \tag{2.287}$$

Otherwise the function has a discontinuity at $x = a$, $y = b$. If a function is defined and continuous at every point of a connected domain, it is called continuous on this domain. Similarly can be defined the continuity of functions of more than two variables.

2.18.5 Properties of Continuous Functions

2.18.5.1 Theorem on Zeros of Bolzano

If a function $f(x, y)$ is defined and continuous in a connected domain, and at two points (x_1, y_1) and (x_2, y_2) of this domain the values have different signs, then there exists at least one point (x_3, y_3) in this domain such that $f(x, y)$ is equal to zero there:

$$f(x_3, y_3) = 0, \quad \text{if} \quad f(x_1, y_1) > 0 \quad \text{and} \quad f(x_2, y_2) < 0\,. \tag{2.288}$$

2.18.5.2 Intermediate Value Theorem

If a function $f(x, y)$ is defined and continuous in a connected domain, and at two points (x_1, y_1) and (x_2, y_2) it has different values $A = f(x_1, y_1)$ and $B = f(x_2, y_2)$, then for an arbitrary value C between A and B there is at least one point (x_3, y_3) such that:

$$f(x_3, y_3) = C, \quad A < C < B \quad \text{or} \quad B < C < A\,. \tag{2.289}$$

2.18.5.3 Theorem About the Boundedness of a Function

If a function $f(x, y)$ is continuous on a bounded and closed domain, it is bounded in this domain, i.e., there are two numbers m and M such that for every point (x, y) in this domain:

$$m \leq f(x, y) \leq M. \tag{2.290}$$

2.18.5.4 Weierstrass Theorem (About the Existence of Maximum and Minimum)

If a function $f(x, y)$ is continuous on a bounded and closed domain, then it takes its maximum and minimum here, i.e., there is at least one point (x', y') such that all the values $f(x, y)$ in this domain are less than or equal to the value $f(x', y')$, and there is at least one point (x'', y'') such that all the values $f(x, y)$ in this domain are greater than or equal to $f(x'', y'')$: For any point (x, y) of this domain

$$f(x', y') \geq f(x, y) \geq f(x'', y'') \qquad (2.291)$$

is valid.

2.19 Nomography

2.19.1 Nomograms

Nomograms are graphical representations of a functional correspondence between three or more variables. From the nomogram, the corresponding values of the variables of a given formula – the *key formula* – in a given domain of the variables can be immediately read directly. Important examples of nomograms are *net charts* and *alignment charts*.
Nomograms are still used in laboratories, even in the computer age, for instance to get approximate values or starting guesses for iterations.

2.19.2 Net Charts

To represent a correspondence between the variables x, y, z given by the equation

$$F(x, y, z) = 0 \qquad (2.292)$$

(or in many cases explicitly by $z = f(x, y)$), the variables can be considered as coordinates in space. The equation (2.292) defines a surface which can be visualized on two-dimensional paper by its level curves (see 2.18.1.2, p. 118). Here, a family of curves is assigned to each variable. These curves form a net: The variables x and y are represented by lines parallel to the axis, the variable z is represented by the family of level curves.

■ Ohm's law is $U = R \cdot I$. The voltage U can be represented by its level curves depending on two variables. If R and I are chosen as Cartesian coordinates, then the equation $U =$ const for every constant corresponds to a hyperbola **(Fig. 2.108)**. By looking at the figure one can tell the corresponding value of U for every pair of values R and I, and also I corresponding to every R, U, and also R corresponding to every I and U. Of course, the investigation is always to be restricted to the domain which is interpreted: In **Fig. 2.108** holds $0 < R < 10$, $0 < I < 10$ and $0 < U < 100$.

Remarks:

1. By changing the calibration, the nomogram can be used for other domains. If, e.g., in **(Fig. 2.108)** the domain $0 < I < 1$ should be represented but R should remain the same, then the hyperbolas of U are to be marked by $U/10$.

2. By application of scales (see 2.17.1, p. 115) it is possible to transform nomograms with complicated curves into straight-line nomograms. Using uniform scales on the x and y axis, every equation of the form

$$x\varphi(z) + y\psi(z) + \chi(z) = 0 \qquad (2.293)$$

can be represented by a nomogram consisting of straight lines. If function scales $x = f(z_2)$ and $y = g(z_2)$ are used, the equation of the form

$$f(z_2)\varphi(z_1) + g(z_2)\psi(z_1) + \chi(z_1) = 0 \qquad (2.294)$$

has a representation for the variables z_1, z_2 and z_3 as two families of curves parallel to the axis and an arbitrary family of straight lines.

■ By applying a logarithmic scale (see 2.17.1, p. 115), Ohm's law can be represented by a straight-line nomogram. Taking the logarithm of $R \cdot I = U$ gives $\log R + \log I = \log U$. Substituting $x = \log R$

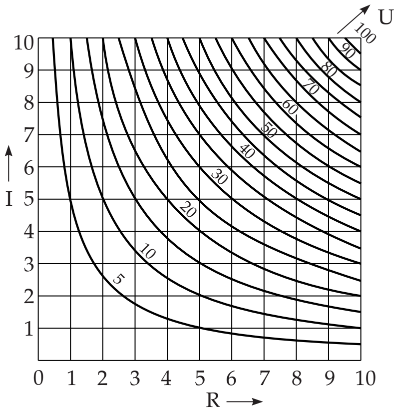


Figure 2.108

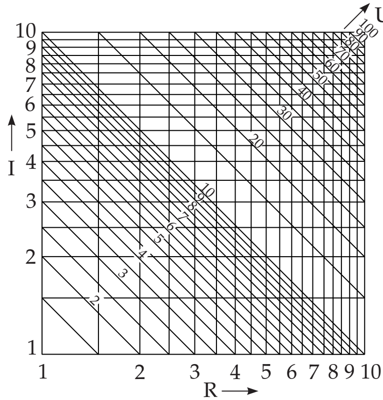


Figure 2.109

and $y = \log I$ results in $x + y = \log U$, i.e., a special form of (2.294). The corresponding nomogram is shown in **Fig. 2.109**.

2.19.3 Alignment Charts

A graphical representation of a relation between three variables z_1, z_2 and z_3 can be given by assigning a scale (see 2.17.1, p. 115) to each variable. The z_i scale has the equation

$$x_i = \varphi_i(z_i), y_i = \psi_i(z_i) \quad (i = 1, 2, 3). \tag{2.295}$$

The functions φ_i and ψ_i are chosen in such a manner that the values of the three variables z_1, z_2 and z_3 satisfying the nomogram equation should lie on a straight line. To satisfy this condition, the area of the triangle, given by the points $(x_1, y_1), (x_2, y_2)$ and (x_3, y_3) , must be zero (see (3.301), p. 195), i.e.,

$$\begin{vmatrix} x_1 & y_1 & 1 \\ x_2 & y_2 & 1 \\ x_3 & y_3 & 1 \end{vmatrix} = \begin{vmatrix} \varphi_1(z_1) & \psi_1(z_1) & 1 \\ \varphi_2(z_2) & \psi_2(z_2) & 1 \\ \varphi_3(z_3) & \psi_3(z_3) & 1 \end{vmatrix} = 0 \tag{2.296}$$

must hold. Every relation between three variables z_1, z_2 and z_3 , which can be transformed into the form (2.296), can be represented by an *alignment nomogram* .
Next, follows the description of some important special cases of (2.296).

2.19.3.1 Alignment Charts with Three Straight-Line Scales Through a Point

If the zero point is chosen for the common point of the lines having the three scales z_1, z_2 or z_3 , then (2.296) has the form

$$\begin{vmatrix} \varphi_1(z_1) & m_1\varphi_1(z_1) & 1 \\ \varphi_2(z_2) & m_2\varphi_2(z_2) & 1 \\ \varphi_3(z_3) & m_3\varphi_3(z_3) & 1 \end{vmatrix} = 0\,, \tag{2.297}$$

since the equation of a line passing through the origin has the equation $y = mx$. Evaluating the determinant (2.297), yields

$$\frac{m_2 - m_3}{\varphi_1(z_1)} + \frac{m_3 - m_1}{\varphi_2(z_2)} + \frac{m_1 - m_2}{\varphi_3(z_3)} = 0 \tag{2.298a}$$

or

$$\frac{C_1}{\varphi_1(z_1)} + \frac{C_2}{\varphi_2(z_2)} + \frac{C_3}{\varphi_3(z_3)} = 0 \quad \text{with } C_1 + C_2 + C_3 = 0\,. \tag{2.298b}$$

Here C_1, C_2 and C_3 are constants.

■ The equation $\frac{1}{a} + \frac{1}{b} = \frac{2}{f}$ is a special case of (2.298b) and it is an important relation, for instance in optics or for the parallel connection of resistances. The corresponding alignment nomogram consists of three uniformly scaled lines.

2.19.3.2 Alignment Charts with Two Parallel Inclined Straight-Line Scales and One Inclined Straight-Line Scale

One of the scales is put on the y-axis, the other one on another line parallel to it at a distance d. The third scale is put on a line $y = mx$. In this case (2.296) has the form

$$\begin{vmatrix} 0 & \psi_1(z_1) & 1 \\ d & \psi_2(z_2) & 1 \\ \varphi_3(z_3) & m\varphi_3(z_3) & 1 \end{vmatrix} = 0. \tag{2.299}$$

Evaluation of the determinant by the first column yields

$$d\,(m\varphi_3(z_3) - \psi_1(z_1)) + \varphi_3(z_3)\,(\psi_1(z_1) - \psi_2(z_2)) = 0. \tag{2.300a}$$

Consequently:

$$\psi_1(z_1)\,\frac{\varphi_3(z_3) - d}{\varphi_3(z_3)} - (\psi_2(z_2) - md) = 0 \ \text{ oder } \ f(z_1)\cdot g(z_3) - h(z_2) = 0. \tag{2.300b}$$

It is often useful to introduce measure scales E_1 and E_2 of the form

$$E_1 f(z_1)\frac{E_2}{E_1} g(z_3) - E_2 h(z_2) = 0. \tag{2.300c}$$

Then, $\varphi_3(z_3) = \dfrac{d}{1 - \dfrac{E_2}{E_1} g(z_3)}$ holds. The relation $E_2 : E_1$ can be chosen so that the third scale is pulled near a certain point or it is gathered. Substituting $m = 0$, then $E_2 h(z_2) = \psi_2(z_2)$ and in this case, the line of the third scale passes through both the starting points of the first and of the second scale. Consequently, these two scales must be placed with a scale division in opposite directions, while the third one will be between them.

■ The relation between the Cartesian coordinates x and y of a point in the x, y plane and the corresponding angle φ in polar coordinates is:

$$y^2 = x^2 \tan^2 \varphi\,. \tag{2.301}$$

The corresponding nomogram is shown in **Fig. 2.110**. The scale division is the same for the scales of x and y but they are oriented in opposite directions. In order to get a better intersection with the third scale between them, their initial points are shifted by a suitable amount. The intersection points of the third scale with the first or with the second one are marked by $\varphi = 0$ or $\varphi = 90°$ respectively.

■ For instance: $x = 3$ $y = 3.5$, gives $\varphi \approx 49.5°$.

2.19.3.3 Alignment Charts with Two Parallel Straight Lines and a Curved Scale

If one of the straight-line scales is placed on the y-axis and the other one is placed at a distance d from it, then equation (2.296) has the form

$$\begin{vmatrix} 0 & \psi_1(z_1) & 1 \\ d & \psi_2(z_2) & 1 \\ \varphi_3(z_3) & \psi_3(z_3) & 1 \end{vmatrix} = 0\,. \tag{2.302}$$

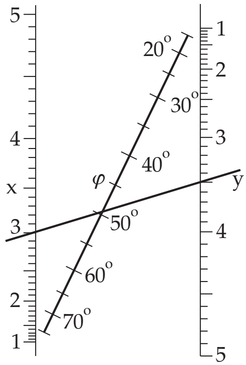


Figure 2.110

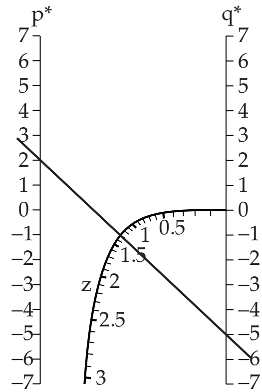


Figure 2.111

Consequently:

$$\psi_1(z_1) + \psi_2(z_2)\frac{\varphi_3(z_3)}{d - \varphi_3(z_3)} - d\frac{\psi_3(z_3)}{d - \varphi_3(z_3)} = 0. \tag{2.303a}$$

Choosing the scale E_1 for the first scale and E_2 for the second one, then (2.303a) is transformed into

$$E_1 f(z_1) + E_2 g(z_2)\frac{E_1}{E_2}h(z_3) + E_1 k(z_3) = 0, \tag{2.303b}$$

where $\psi_1(z_1) = E_1 f(z_1)$, $\psi_2(z_2) = E_2 g(z_2)$ and

$$\varphi_3(z_3) = \frac{dE_1 h(z_3)}{E_2 + E_1 h(z_3)} \quad \text{and} \quad \psi_3(z_3) = -\frac{E_1 E_2 k(z_3)}{E_2 + E_1 h(z_3)} \tag{2.303c}$$

holds.

■ The reduced third-degree equation $z^3 + p^* z + q^* = 0$ (see 1.6.2.3, p. 40) is of the form (2.303b). After the substitutions $E_1 = E_2 = 1$ and $f(z_1) = q^*$, $g(z_2) = p^*$, $h(z_3) = z$, the formulas to calculate the coordinates of the curved scale are $x = \varphi_3(z) = \dfrac{d \cdot z}{1+z}$ and $y = \psi_3(z) = -\dfrac{z^3}{1+z}$.

In **Fig. 2.111** the curved scale is shown only for positive values of z. The negative values one gets by replacing z by $-z$ and by the determination of the positive roots from the equation $z^3 + p^* z - q^* = 0$. The complex roots $u + iv$ can also be determined by nomograms. Denoting the real root, which always exists, by z_1, then the real part of the complex root is $u = -z_1/2$, and the imaginary part v can be determined from the equation $3u^2 - v^2 + p^* = \frac{3}{4}z_1^2 - v^2 + p^* = 0$.

■ For instance: $y^3 + 2y - 5 = 0$, i.e., $p^* = 2, q^* = -5$. One reads $z_1 \approx 1.3$.

2.19.4 Net Charts for More Than Three Variables

To construct a chart for formulas containing more than three variables, the expression is to decompose by the help of auxiliary variables into several formulas, each containing only three variables. Here, every auxiliary variable must be contained in exactly two of the new equations. Each of these equations is to be represented by an alignment chart so that the common auxiliary variable has the same scale.

3 Geometry

3.1 Plane Geometry

3.1.1 Basic Notations

3.1.1.1 Point, Line, Ray, Segment

1. Point and Line

Points and *straight lines* are not defined in today's mathematics. The relations between them are determined only by axioms. The *line* is graphically imaginable as a trace of a point moving in a plane along the shortest route between two different points without changing its direction.

A *point* is the intersection of two lines.

2. Closed Half-Line or Ray, and Segment

A *ray* is the set of points of a line which are exactly on one side of a given point O, including this point O. A ray is imaginable as the trace of a point which starts at O and moves along the line without changing its direction, like a beam of light after its emission until it is not led out of its way.

A *segment* $\overline{AB}$ is the set of points of a line lying between two given points A and B of this line, including the points A and B. The segment is the shortest connection between the two points A and B in a plane. The direction class of a segment is denoted by an arrowhead $\overrightarrow{AB}$, or its direction starts at the first mentioned point A, and ends at the second B.

3. Parallel and Orthogonal Lines

Parallel lines run in the same direction; they have no common points, i.e., they do not move off and do not approach each other, and they do not have any intersection point. The *parallelism* of two lines g and g' is denoted by $g||g'$.

Orthogonal lines form a right angle at their intersection, i.e., they are perpendicular to each other. *Orthogonality* and parallelism are mutual positions of two lines.

3.1.1.2 Angle

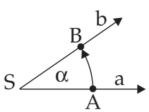


Figure 3.1

1. Notion of Angle

An *angle* is defined by two rays a and b starting at the same point S, so they can be transformed into each other by a rotation **(Fig. 3.1)**. If A is a point on the ray a and B is on the ray b, then the angle in the direction given in **Fig. 3.1** is denoted by the symbols (a, b) or by $\sphericalangle ASB$, or by a Greek letter. The point S is called the *vertex*, the rays a and b are called *sides* or *legs* of the angle.

In mathematics, an angle is called positive or negative depending on the rotation being counterclockwise or clockwise respectively. It is important to distinguish the angle $\sphericalangle ASB$ from the angle $\sphericalangle BSA$. Actually, $\sphericalangle ASB = -\sphericalangle BSA$ $(0 \le \sphericalangle ASB \le 180°)$ and $\sphericalangle ASB = 360° - \sphericalangle BSA$ $(180° \le \sphericalangle ASB \le 360°)$ holds.

Remark: In geodesy the positive direction of rotation is defined by the clockwise direction (see 3.2.2.1, p. 144).

2. Names of Angles

Angles have different names according to the different positions of their legs. The names given in **Table 3.1** are used for angles α in the interval $0 \le \alpha \le 360°$ (see also **Fig. 3.2**).

I.N. Bronshtein et al., *Handbook of Mathematics*,
DOI 10.1007/978-3-662-46221-8_3

Table 3.1 Names of angles in degree and radian measure

Names of angles	Degree	Radian	Names of angles	Degree	Radian
round (full) angle	$\alpha^\circ = 360^\circ$	$\alpha = 2\pi$	right angle	$\alpha^\circ = 90^\circ$	$\alpha = \pi/2$
convex angle	$\alpha^\circ > 180^\circ$	$\pi < \alpha < 2\pi$	acute angle	$0^\circ < \alpha^\circ < 90^\circ$	$0^\circ < \alpha < \pi/2$
straight angle	$\alpha = 180^\circ$	$\alpha = \pi$	obtuse angle	$90^\circ < \alpha < 180^\circ$	$\pi/2 < \alpha < \pi$

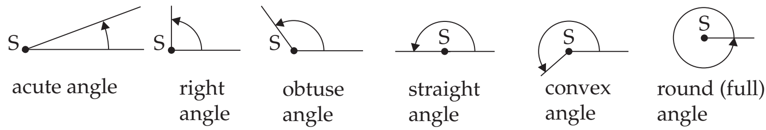


Figure 3.2

3.1.1.3 Angle Between Two Intersecting Lines

At the intersection point of two lines g_1, g_2 there are four angles $\alpha, \beta, \gamma, \delta$ **(Fig. 3.3)**. There are to be distinguished *adjacent angles*, *vertex angles*, *complementary angles*, and *supplementary angles*.

1. Adjacent Angles are neighboring angles at the intersection point of two lines with a common vertex S, and with a common leg; both non-common legs are on the same line, they are rays starting from S but in opposite directions, so adjacent angles sum to 180°.

■ In **Fig. 3.3** the pairs (α, β), (β, γ), (γ, δ) and (α, δ) are adjacent angles.

2. Vertex Angles are the angles at the intersection point of two lines, opposite to each other, having the same vertex S but no common leg, and being equal. They sum to 180° by the same adjacent angle.

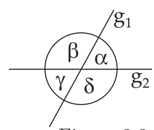


Figure 3.3

■ In **Fig. 3.3** (α, γ) and (β, δ) are vertex angles.

3. Complementary Angles are two angles summing to 90°.

4. Supplementary Angles are two angles summing to 180°.

■ In **Fig. 3.3** the pairs of angles (α, β) or (γ, δ) are supplementary angles.

3.1.1.4 Pairs of Angles with Intersecting Parallels

Intersecting two parallel lines p_1, p_2 by a third one g yields eight angles **(Fig. 3.4)**. Besides the adjacent and vertex angles with the same vertex S , alternate angles, opposite angles, and corresponding (exterior–interior) angles are to be distinguished with different vertices.

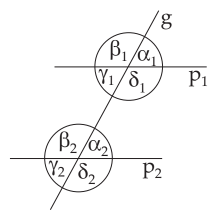


Figure 3.4

1. Alternate Angles have the same size. They are on the opposite sides of the intersecting line g and of the parallel lines p_1, p_2. The legs of alternate angles are in pairs oppositely oriented.

■ In **Fig.3.4**, e.g., (α_1, γ_2), (β_1, δ_2), (γ_1, α_2), (δ_1, β_2) are alternate angles.

2. Corresponding or Exterior–Interior Angles have the same size. They are on the same sides of the intersecting line g and of the parallel lines p_1, p_2. The legs of corresponding angles are oriented in pairs in the same direction.

■ In **Fig. 3.4** the pairs of angles (α_1, α_2), (β_1, β_2), (γ_1, γ_2), and (δ_1, δ_2) are corresponding angles.

3. Opposite Angles are on the same side of the intersecting line g but on different sides of the parallel lines p_1, p_2. They sum to 180°. One pair of legs has the same orientation, the other one is oppositely oriented.

■ In **Fig. 3.4**, e.g., the pairs of angles (α_1, δ_2), (β_1, γ_2), (γ_1, β_2), and (δ_1, α_2) are opposite angles.

3.1.1.5 Angles Measured in Degrees and in Radians

In geometry, the measurement of angles is based on the division of the *full angle* into 360 equal parts or 360° (degrees). This is called *measure in degrees.* Further division of degrees is not a decimal one, it is sexagesimal: $1° = 60'$ (minute, or sexagesimal minute), $1' = 60''$ (second, or sexagesimal second). For grade measure see 3.2.2.2, p. 146 and the remark below.
Besides measure in degrees the radian measure is used to define the size of an angle. The size of the *central angle* α of an arbitrary circle **(Fig. 3.5a)** is given as the ratio of the corresponding arc-length l and the radius of the circle r:

$$\alpha = \frac{l}{r}. \tag{3.1}$$

1 rad	$= 57°\,17'\,44.8'' = 57.2958°$,
1°	$= 0.017453$ rad,
1′	$= 0.000291$ rad,
1″	$= 0.000005$ rad.

The *unit of radian measure* is the *radian* (rad), i.e., the central angle belonging to an arc with arclength l equal to the radius r. In the table you will find approximate conversion values.
If the measure of the angle is $\alpha°$ in degrees and α in radian measure, then for conversion

$$\alpha° = \varrho\alpha = 180°\frac{\alpha}{\pi}, \qquad \alpha = \frac{\alpha°}{\varrho} = \frac{\pi}{180°}\alpha° \quad \text{with} \quad \varrho = \frac{180°}{\pi}. \tag{3.2}$$

holds. In particular: $360° = 2\pi$, $180° = \pi$, $90° = \pi/2$, $270° = 3\pi/2$, etc. Formulas (3.2) refer to decimal fractions and the following examples show how to make calculations with minutes and seconds.

■ **A:** Conversion of an angle given in degrees into radian measure:
$52°\,37'\,23'' = 52 \cdot 0.017453 + 37 \cdot 0.000291 + 23 \cdot 0.000005 = 0.918447\,\text{rad}$.

■ **B:** Conversion of an angle given in radians into degrees:
$5.645\,\text{rad} = 323 \cdot 0.017453 + 26 \cdot 0.000291 + 5 \cdot 0.000005 = 323°\,26'\,05''$.
The result is getting from:
$5.645 : 0.017453 = 323 + 0.007611$
$0.007611 : 0.000291 = 26 + 0.000025$
$0.000025 : 0.000005 = 5$.
The notation rad is usually omitted if it is obvious from the text that the number refers to the radian measure of an angle (see also p. 1055).

Remark: In geodesy a full angle is divided into 400 equal parts, called grades. This is called *measure in grades.* A right angle is 100 gon. A gon is divided into 1000 mgon.
On calculators the notation DEG is used for degree, GRAD for grade, and RAD for radian. For conversion between the different measures see Table 3.5, p. 146.

3.1.2 Geometrical Definition of Circular and Hyperbolic Functions

3.1.2.1 Definition of Circular or Trigonometric Functions

1. Definition by the Unit Circle

The trigonometric functions of an angle α are defined for both the unit circle with radius $R = 1$ and the acute angles of a right-angled triangle **(Fig. 3.5a,b)** with the help of the *adjacent side* b, *opposite side* a, and *hypotenuse* c. In the unit circle the measurement of an angle is made between a fixed radius $\overline{OA}$ (with length 1) and a moving radius $\overline{OC}$ counterclockwise (positive direction):

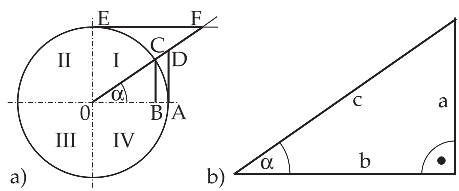


Figure 3.5

sine : $\sin\alpha = \overline{BC} = \dfrac{a}{c}$, (3.3) *cosine* : $\cos\alpha = \overline{OB} = \dfrac{b}{c}$, (3.4)

tangent : $\tan\alpha = \overline{AD} = \dfrac{a}{b}$, (3.5) *cotangent* : $\cot\alpha = \overline{EF} = \dfrac{b}{a}$, (3.6)

secant : $\sec\alpha = \overline{OD} = \dfrac{c}{b}$, (3.7) *cosecant* : $\csc\alpha = \overline{OF} = \dfrac{c}{a}$. (3.8)

2. Signs of Trigonometric Functions

Depending in which quadrant of the unit circle **(Fig. 3.5a)** the moving radius $\overline{OC}$ is, the functions have well-defined signs which can be taken from **Table 2.2** (see p. 79).

3. Definition of Trigonometric Functions by Area of Circular Sectors

The functions $\sin\alpha$, $\cos\alpha$, $\tan\alpha$, $\cot\alpha$ are defined by the segments $\overline{BC}$, $\overline{OB}$, $\overline{AD}$ of the unit circle with $R = 1$ **(Fig. 3.6)**, where the argument is the central angle $\alpha = \sphericalangle AOC$. For this definition one could use also the area x of the sector COK, which is denoted in **Fig. 3.6** by the shaded area. With the central angle 2α measured in radians, one gets $x = \dfrac{1}{2}R^2 2\alpha = \alpha$ for the area with $R = 1$. Therefore, one has the same equations for $\sin x = \overline{BC}$, $\cos x = \overline{OB}$, $\tan x = \overline{AD}$ as in (3.3, 3.4, 3.5).

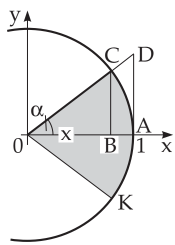


Figure 3.6

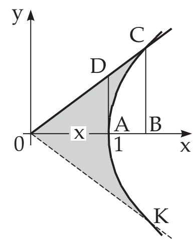


Figure 3.7

3.1.2.2 Definitions of the Hyperbolic Functions

In analogy with the definition of the trigonometric functions in (3.3), (3.4), (3.5) now instead of the area of a sector of the unit circle with the equation $x^2 + y^2 = 1$ here the corresponding area of the sector of the hyperbola is used with the equation $x^2 - y^2 = 1$ (only the right branch in **Fig. 3.7**). Denoting by x the area of COK, the shaded area in **Fig. 3.7**, the defining equations of the hyperbolic functions are:

$\sinh x = \overline{BC}$, (3.9) $\cosh x = \overline{OB}$, (3.10) $\tanh x = \overline{AD}$. (3.11)

Calculating the area x by integration and expressing the results in terms of $\overline{BC}$, $\overline{OB}$, and $\overline{AD}$ yields

$$x = \ln(\overline{BC} + \sqrt{\overline{BC}^2 + 1}) = \ln(\overline{OB} + \sqrt{\overline{OB}^2 - 1}) = \frac{1}{2}\ln\frac{1 + \overline{AD}}{1 - \overline{AD}}, \tag{3.12}$$

and so, from now on, the hyperbolic functions can be expressed in terms of exponential functions:

$$\overline{BC} = \frac{e^x - e^{-x}}{2} = \sinh x, \tag{3.13a}$$

$$\overline{OB} = \frac{e^x + e^{-x}}{2} = \cosh x, \tag{3.13b}$$

$$\overline{AD} = \frac{e^x - e^{-x}}{e^x + e^{-x}} = \tanh x. \tag{3.13c}$$

These equations represent the most popular definition of the hyperbolic functions.

3.1.3 Plane Triangles

3.1.3.1 Statements about Plane Triangles

1. The Sum of Two Sides of a plane triangle is greater than the third one **(Fig. 3.8)**:

$$b + c > a. \tag{3.14}$$

2. The Sum of the Angles of a plane triangle is

$$\alpha + \beta + \gamma = 180^\circ . \qquad (3.15)$$

3. Unique Determination of Triangles A triangle is uniquely determined by the following data:
- by three sides or
- by two sides and the angle between them or
- by one side and the two angles on it.

If two sides and the angle opposite one of them are given, then they define two, one or no triangle (see the third basic problem in **Table 3.4**, p. 145).

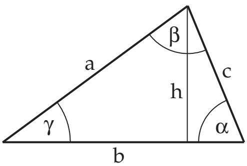


Figure 3.8

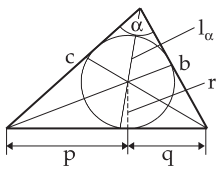


Figure 3.9

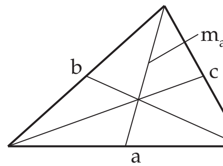


Figure 3.10

4. Median of a Triangle is a line connecting a vertex of the triangle with the midpoint of the opposite side. The medians of the triangle intersect each other at one point, at the *center of gravity* of the triangle **(Fig. 3.10)**, which divides them in the ratio 2 : 1 counting from the vertex.

5. Bisector of a Triangle is a line which divides one of the interior angles into two equal parts. The bisectors intersect each other at one point.

6. Incircle is the circle inscribed in a triangle, i.e., all the sides are tangents of the circle. Its center is the intersection point of the bisectors **(Fig. 3.9)**. The radius of the inscribed circle is called the *apothem* or the *short radius*.

7. Circumcircle is the circle drawn around a triangle, i.e., passing through the vertices of the triangle **(Fig. 3.11)**. Its center is the intersection point of the three *right bisectors* of the triangle.

8. Altitude of a Triangle is the perpendicular line that starts at a vertex and is perpendicular to the opposite side. The altitudes intersect each other at one point, the *orthocenter*.

9. Isosceles Triangle In an *isosceles triangle* two sides have equal length. The altitude, median, and bisector of the third side coincide. For a triangle the equality of any two of these sides is enough to make it isosceles.

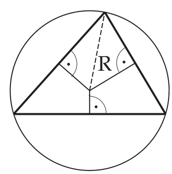


Figure 3.11

10. Equilateral Triangle In an *equilateral triangle* with $a = b = c$ the centers of the incircle and the circumcircle, the center of gravity, and the orthocenter coincide.

11. Median is a line connecting two midpoints of sides of a triangle; it is parallel to the third side and has the half of length as that side.

12. Right-Angled Triangle is a triangle that has a right angle (an angle of 90°) (**Fig. 3.31**, see p. 142).

3.1.3.2 Symmetry

1. Central Symmetry

A plane figure is called *center symmetric* if by a rotation of the plane by 180° around the *central point* or the *center of symmetry S* it exactly covers itself **(Fig. 3.12)**. Because the size and shape of the figure do not change during this transformation, it is called a *congruent mapping*. Also the *sense class* or *orientation class* of the plane figure remains the same **(Fig. 3.12)**. Because of the same sense class such figures are called *directly congruent*.

The *orientation of a figure* means the traverse of the boundary of a figure in a direction: positive direction, hence counterclockwise, negative direction, hence clockwise (**Fig. 3.12, Fig. 3.13**).

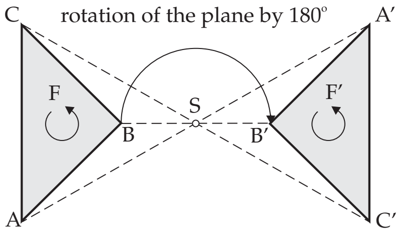


Figure 3.12

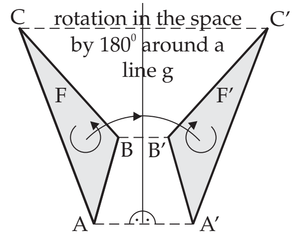


Figure 3.13

2. Axial Symmetry

A plane figure is called *axially symmetric* if the corresponding points cover each other after a rotation in space by 180° around a line g **(Fig. 3.13)**. The corresponding points have the same distance from the axis g, the axis of symmetry. The orientation of the figure is reversed for axial symmetry with respect to the line g. Therefore, such figures are called *indirectly congruent*. This transformation is called a *reflection in g*. Because the size and the shape of the figures do not change, it is also called a *indirect congruent mapping*. The *orientation class* of the plane figure is reversed under this transformation **(Fig. 3.13)**.

Remark: For space figures hold the analogous statements.

3. Congruent Triangles, Congruence Theorems

a) Congruence: Plane figures are called *congruent* if their size and shape coincide. Congruent figures can be transformed into a position of superimposition by the following three transformations, by *translation*, *rotation*, and *reflection*, and combinations of these.
It is to distinguish between *directly congruent figures* and *indirectly congruent figures*. Directly congruent figures can be transformed into a covering position by translation and rotation. Because the indirectly congruent figures have a reversed sense class, an axially symmetric transformation with respect to a line is also needed to transform them into a covering position.

■ Axially symmetric figures are indirectly congruent. To transform them into each other all three transformations are needed.

b) Laws of Congruence for Triangles: Two triangles are congruent if they coincide for:
- three sides (SSS) or
- two sides and the angle between them (SAS), or
- one side and the interior angles on this side (ASA), or
- two sides and the interior angle being opposite to the longer one (SSA).

4. Similar Triangles, Similarity Theorems

Plane figures are called *similar* if they have the same shape without having the same size. For similar figures there is a one-to-one mapping between their points such that every angle in one figure is the same as the corresponding angle in the other figure. An equivalent definition is the following: In similar figures the length of segments corresponding to each other are proportional.

a) Similarity of figures requires either the equality of all the corresponding angles or the equality of the ratio of all corresponding segments.

b) Area The *areas of similar plane figures* are proportional to the square of the ratio of corresponding linear elements such as sides, altitudes, diagonals, etc.

c) Laws of Similarity For triangles the following laws of similarity hold. Triangles are similar if they coincide for:

- two ratios of sides,
- two interior angles,
- the ratio of two sides and the interior angle between them,
- the ratio of two sides and the interior angle opposite of the longer one.

Because in the laws of similarity only the equality of ratios of sides is required and not the equality of length of sides, therefore the laws of similarity require less than the corresponding laws of congruence.

5. Intercept Theorems

The *intercept theorems* are a consequence of the laws of similarity of a triangle.

1. First Intercept Theorem If two rays starting at the same point S are intersected by two parallels p_1, p_2, then the segments of one of the rays **(Fig. 3.14a)** have the same ratio as the corresponding segments on the other one:

$$\left|\frac{\overline{SP_1}}{\overline{SQ_1}}\right| = \left|\frac{\overline{SP_2}}{\overline{SQ_2}}\right|. \qquad (3.16)$$

Consequently, every segment on one of the rays is proportional to the corresponding segment on the other ray.

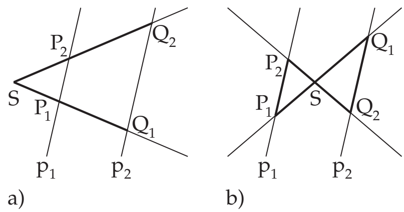


Figure 3.14

2. Second Intercept Theorem If two rays starting at the same point S are intersected by two parallels p_1, p_2 , then the segments of the parallels have the same ratio as the corresponding segments on the rays **(Fig. 3.14a)**:

$$\left|\frac{\overline{SP_1}}{\overline{SQ_1}}\right| = \left|\frac{\overline{P_1P_2}}{\overline{Q_1Q_2}}\right| \quad \text{or} \quad \left|\frac{\overline{SP_2}}{\overline{SQ_2}}\right| = \left|\frac{\overline{P_1P_2}}{\overline{Q_1Q_2}}\right|. \qquad (3.17)$$

The intercept theorems are also valid in the case of intersecting lines at the point S, if the point S is between the parallels **(Fig. 3.14b)**.

3.1.4 Plane Quadrangles

3.1.4.1 Parallelogram

A quadrangle is called a parallelogram **(Fig. 3.15)** if it has the following properties:

- the sides opposite to each other have the same length,
- the sides opposite to each other are parallel,
- the diagonals intersect each other at their midpoints,
- the angles opposite to each other are equal.

Supposing only one of the previous properties for a quadrangle, or supposing the equality and the parallelism of one pair of opposite sides, then all the other properties follow from it.
The relations between diagonals, sides, and area are the following:

$$d_1^2 + d_2^2 = 2(a^2 + b^2), \quad (3.18) \qquad h = b \sin\alpha, \quad (3.19) \qquad S = ah. \quad (3.20)$$

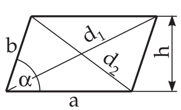


Figure 3.15

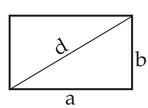


Figure 3.16

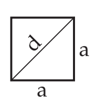


Figure 3.17

3.1.4.2 Rectangle and Square

A parallelogram is a *rectangle* **(Fig. 3.16)**, if it has:
- only right angles, or
- the diagonals have the same length.

Only one of these properties is enough, because either of them follows from the other. It is sufficient to show that one angle of the parallelogram is a right angle, then all the angles are right angles. If a quadrangle has four right angles, it is a rectangle.
The perimeter U and the area S of a rectangle are:

$$U = 2(a+b), \quad (3.21a) \qquad S = ab. \quad (3.21b)$$

If $a = b$ holds **(Fig. 3.17)**, the rectangle is called a *square*, and the following formulas

$$d = a\sqrt{2} \approx 1.414a, \quad (3.22) \qquad a = d\frac{\sqrt{2}}{2} \approx 0.707d, \quad (3.23) \qquad S = a^2 = \frac{d^2}{2}. \quad (3.24)$$

3.1.4.3 Rhombus

A rhombus **(Fig. 3.18)** is a parallelogram in which
- all the sides have the same length, or
- the diagonals are perpendicular to each other, or
- the diagonals are bisectors of the angles of the parallelogram.

Any of the previous properties is enough alone; all the others follow from it. For the rhombus

$$d_1 = 2a\cos\frac{\alpha}{2}, \quad (3.25) \qquad d_2 = 2a\sin\frac{\alpha}{2}, \quad (3.26) \qquad d_1^2 + d_2^2 = 4a^2. \quad (3.27)$$

$$S = ah = a^2\sin\alpha = \frac{d_1 d_2}{2}. \quad (3.28)$$

3.1.4.4 Trapezoid

A quadrangle is called trapezoid if it has two parallel sides **(Fig. 3.19)**. The parallel sides are called *bases*. With the notation a and b for the bases, h for the *altitude* and m for the *median of the trapezoid* which connects the midpoints of the two non-parallel sides

$$m = \frac{a+b}{2}, \quad (3.29) \qquad S = \frac{(a+b)h}{2} = mh, \quad (3.30) \qquad h_S = \frac{h(a+2b)}{3(a+b)}. \quad (3.31)$$

The centroid is on the connecting segment of the midpoints of the parallel basis a and b, at a distance h_S (3.31) from the base a. For the calculation of the coordinates of the centroid by integration see 8.2.2.3, p. 506.
For an *isosceles trapezoid* with $d = c$

$$S = (a - c\cos\gamma)c\sin\gamma = (b + c\cos\gamma)c\sin\gamma. \quad (3.32)$$

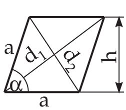


Figure 3.18

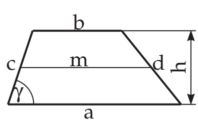


Figure 3.19

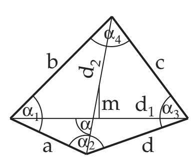


Figure 3.20

3.1.4.5 General Quadrangle

A closed plane figure bounded by four straight line segments is called a *general quadrangle*. If the diagonals lie fully inside the quadrangle, it is *convex*, otherwise *concave*. The general quadrangle is

divisible by two diagonals d_1, d_2 in two triangles (**Fig. 3.20**). Therefore, in every quadrangle the sum of the interior angles is $2 \cdot 180° = 360°$:

$$\sum_{i=1}^{4} \alpha_i = 360°. \tag{3.33}$$

The length of the segment m connecting the midpoints of the diagonals **(Fig. 3.20)** is given by

$$a^2 + b^2 + c^2 + d^2 = d_1^2 + d_2^2 + 4m^2. \tag{3.34}$$

The area of the general quadrangle is

$$S = \frac{1}{2} d_1 d_2 \sin\alpha. \tag{3.35}$$

3.1.4.6 Inscribed Quadrangle

A quadrangle which can be circumscribed by a circumcircle is called an *inscribed quadrangle* **(Fig. 3.21a)** and its sides are chords of this circle. A quadrangle is an *inscribed quadrangle* if and only if the sums of its opposite angles are 180°:

$$\alpha + \gamma = \beta + \delta = 180°. \tag{3.36}$$

Ptolemy's theorem is valid for the insribed quadrangle:

$$ac + bd = d_1 d_2. \tag{3.37}$$

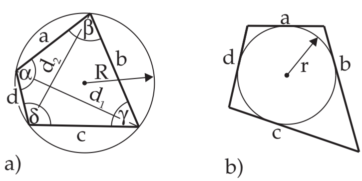


Figure 3.21

The radius of the circumcircle of an inscribed quadrangle is

$$R = \frac{1}{4S}\sqrt{(ab+cd)(ac+bd)(ad+cb)}. \tag{3.38}$$

The diagonals can be calculated by the formulas

$$d_1 = \sqrt{\frac{(ac+bd)(ab+cd)}{ad+bc}}, \tag{3.39a}$$

$$d_2 = \sqrt{\frac{(ac+bd)(ad+bc)}{ab+cd}}. \tag{3.39b}$$

The area can be expressed in terms of the half-perimeter of the quadrangle $s = \frac{1}{2}(a+b+c+d)$:

$$S = \sqrt{(s-a)(s-b)(s-c)(s-d)}. \tag{3.40}$$

If the inscribed quadrangle is also a circumscribing quadrangle (see **Fig. 3.21** and 3.1.4.7), then

$$S = \sqrt{abcd}. \tag{3.41}$$

3.1.4.7 Circumscribing Quadrangle

If a quadrangle has an *inscribed circle* **(Fig. 3.21b)**, then it is called a *circumscribing quadrangle*, and the sides are tangents to the circle. A quadrangle has an inscribed circle if and only if the sum of the lengths of the opposite sides are equal, and this sum is also equal to the half-perimeter s:

$$s = \frac{1}{2}(a+b+c+d) = a+c = b+d. \tag{3.42}$$

The area of the circumscribing quadrangle is

$$S = (a+c)r = (b+d)r, \tag{3.43}$$

where r is the radius of the inscribed circle.

3.1.5 Polygons in the Plane

3.1.5.1 General Polygon

A closed plane figure bounded by straight-line segments as its sides, can be decomposed into $n-2$ triangles **(Fig. 3.22)**. The sums of the exterior angles β_i, and of the interior angles γ_i, and the number of diagonals D are

$$\sum_{i=1}^{n} \beta_i = 360^\circ, \quad (3.44) \qquad \sum_{i=1}^{n} \gamma_i = 180^\circ(n-2), \quad (3.45) \qquad D = \frac{n(n-3)}{2}. \quad (3.46)$$

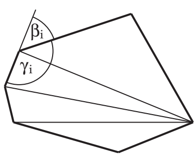
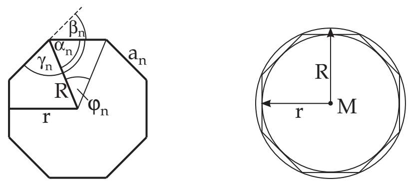


Figure 3.22



a) b)

Figure 3.23

3.1.5.2 Regular Convex Polygons

Regular convex polygons **(Fig. 3.23)** have n equal sides and n equal angles. The intersection point of the mid-perpendiculars of the sides is the center M of the inscribed and of the circumscribed circle with radii r and R, respectively. The sides of these polygons are tangents to the inscribed circle and chords of the the circumcircle. They form a *circumscribing polygon* or *tangent polygon* for the inscribed circle and a inscribed polygon in the circumcircle. The decomposition of a regular convex n gon (regular convex polygon) results in n isosceles congruent triangles around the center M.

Central Angle $\varphi_n = \dfrac{360^\circ}{n}$. (3.47) **Base Angle** $\alpha_n = \left(1 - \dfrac{2}{n}\right) \cdot 90^\circ$. (3.48)

Exterior Angle $\beta_n = \dfrac{360^\circ}{n}$. (3.49) **Interior Angle** $\gamma_n = 180^\circ - \beta_n$. (3.50)

Circumcircle Radius
$$R = \frac{a_n}{2\sin\dfrac{180^\circ}{n}}, \quad R^2 = r^2 + \frac{1}{4}a_n^2. \tag{3.51}$$

Inscribed Circle Radius
$$r = \frac{a_n}{2}\cot\frac{180^\circ}{n} = R\cos\frac{180^\circ}{n}. \tag{3.52}$$

Side Length $a_n = 2\sqrt{R^2 - r^2} = 2R\sin\dfrac{\varphi_n}{2} = 2r\tan\dfrac{\varphi_n}{2}$. (3.53) **Perimeter** $U = na_n$. (3.54)

Side Length of the 2*n* gon
$$a_{2n} = R\sqrt{2 - 2\sqrt{1 - \left(\frac{a_n}{2R}\right)^2}}, \quad a_n = a_{2n}\sqrt{4 - \left(\frac{a_{2n}^2}{R^2}\right)}. \tag{3.55}$$

Area of the *n* gon
$$S_n = \frac{1}{2}na_n r = nr^2\tan\frac{\varphi_n}{2} = \frac{1}{2}nR^2\sin\varphi_n = \frac{1}{4}na_n^2\cot\frac{\varphi_n}{2}. \tag{3.56}$$

Area of the 2n gon $$S_{2n} = \frac{nR^2}{\sqrt{2}}\sqrt{1-\sqrt{\frac{4S_n^2}{n^2R^4}}}\,, \quad S_n = S_{2n}\sqrt{1-\frac{S_{2n}^2}{n^2R^4}}\,. \qquad (3.57)$$

3.1.5.3 Some Regular Convex Polygons

The properties of some regular convex polygons are collected in **Table 3.2**.

The pentagon and the pentagram deserve special attention since it is presumed that Hippasos of Metapontum (ca. 450 BC) recognized irrational numbers by the properties of these polygons (see 1.1.1.2, p. 2). A discussion follows in the example:

■ The diagonals of a regular pentagon **(Fig. 3.24)** form an *inscribed pentagram*. Its sides enclose a regular pentagon again. In a regular pentagon, the proportion of a diagonal and a side is equal to the proportion of a side and the (diagonal minus side): $a_0 : a_1 = a_1 : (a_0 - a_1) = a_1 : a_2$, where $a_2 = a_0 - a_1$.
Considering smaller and smaller nested pentagons with $a_3 = a_1 - a_2, a_4 = a_2 - a_3, \ldots$ and $a_2 < a_1, a_3 < a_2, a_4 < a_3 \ldots$, yields $a_0 : a_1 = a_1 : a_2 = a_2 : a_3 = a_3 : a_4 = \cdots$. The Euclidean algorithm for a_0 and a_1 never breaks off, since $a_0 = 1 \cdot a_1 + a_2, a_1 = 1 \cdot a_2 + a_3, a_2 = 1 \cdot a_3 + a_4, \ldots$, hence $q_n = 1$. The side a_1 and diagonal a_0 of the regular pentagon are incommensurable. The continued fraction determined by $a_0 : a_1$ is identical to the golden section in ■ **B**, 1.1.1.4, **3.**, p. 4, i.e., it results in an irrational number.

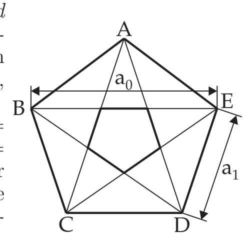


Figure 3.24

3.1.6 The Circle and Related Shapes

3.1.6.1 Circle

The *circle* is the locus of the points in a plane which are at the same given distance from a given point, the *center of the circle*. The distance itself, and also the line segment connecting the center with any point of the circle, is called the *radius*. The *circumference* or *periphery of the circle* encloses the *area of the circle*. A line segment connecting two points of the circle is called a *chord*. A line passing through two points of a circle is called a *secant*. Lines having exactly one common point with the circle are called *tangent lines* or *tangents* of the circle.

Chord Theorem (Fig. 3.26) $$\overline{AC} \cdot \overline{AD} = \overline{AB} \cdot \overline{AE} = r^2 - m^2. \qquad (3.58)$$

Secant Theorem (Fig. 3.27) $$\overline{AB} \cdot \overline{AE} = \overline{AC} \cdot \overline{AD} = m^2 - r^2. \qquad (3.59)$$

Secant–Tangent Theorem (Fig. 3.27) $$\overline{AT}^2 = \overline{AB} \cdot \overline{AE} = \overline{AC} \cdot \overline{AD} = m^2 - r^2. \qquad (3.60)$$

Perimeter $$U = 2\pi r \approx 6,283r, \quad U = \pi d \approx 3,142d, \quad U = 2\sqrt{\pi S} \approx 3,545\sqrt{S}. \qquad (3.61)$$

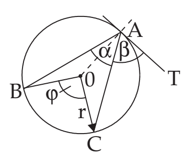


Figure 3.25

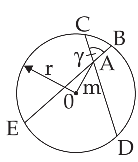


Figure 3.26

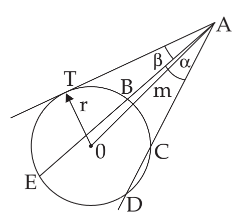


Figure 3.27

Table 3.2 Properties of some regular polygons

	a_n	R	r	S_n
3 gon	$a_3 = R\sqrt{3} = 2r\sqrt{3}$	$= \frac{a_3}{3}\sqrt{3} = 2r = \frac{2}{3}h$ $h = \frac{a_3}{2}\sqrt{3} = \frac{3}{2}R$	$= \frac{a_3}{6}\sqrt{3} = \frac{R}{2} = \frac{1}{3}h$	$= \frac{a_3^2}{4}\sqrt{3} = \frac{3R^2}{4}\sqrt{3}$ $= 3r^2\sqrt{3}$
5 gon	$a_5 = \frac{R}{2}\sqrt{10-2\sqrt{5}}$ $= 2r\sqrt{5-2\sqrt{5}}$	$= \frac{a_5}{10}\sqrt{50+10\sqrt{5}}$ $= r(\sqrt{5}-1)$	$= \frac{a_5}{10}\sqrt{25+10\sqrt{5}}$ $= \frac{R}{4}(\sqrt{5}+1)$	$= \frac{a_5^2}{4}\sqrt{25+10\sqrt{5}}$ $= \frac{5R^2}{8}\sqrt{10+2\sqrt{5}}$ $= 5r^2\sqrt{5-2\sqrt{5}}$
6 gon	$a_6 = \frac{2}{3}r\sqrt{3}$	$= \frac{2}{3}r\sqrt{3}$	$= \frac{R}{2}\sqrt{3}$	$= \frac{3a_6^2}{2}\sqrt{3} = \frac{3R^2}{2}\sqrt{3}$ $= 2r^2\sqrt{3}$
8 gon	$a_8 = R\sqrt{2-\sqrt{2}}$ $= 2r(\sqrt{2}-1)$	$= \frac{a_8}{2}\sqrt{4+2\sqrt{2}}$ $= r\sqrt{4-2\sqrt{2}}$	$= \frac{a_8}{2}(\sqrt{2}+1)$ $= \frac{R}{2}\sqrt{2+\sqrt{2}}$	$= 2a_8^2(\sqrt{2}+1)$ $= 2R^2\sqrt{2}$ $= 8r^2(\sqrt{2}-1)$
10 gon	$a_{10} = \frac{R}{2}(\sqrt{5}-1)$ $= \frac{2r}{5}\sqrt{25-10\sqrt{5}}$	$= \frac{a_{10}}{2}(\sqrt{5}+1)$ $= \frac{r}{5}\sqrt{50-10\sqrt{5}}$	$= \frac{a_{10}}{2}\sqrt{5+2\sqrt{5}}$ $= \frac{R}{4}\sqrt{10+2\sqrt{5}}$	$= \frac{5a_{10}^2}{2}\sqrt{5+2\sqrt{5}}$ $= \frac{5R^2}{4}\sqrt{10-2\sqrt{5}}$ $= 2r^2\sqrt{25-10\sqrt{5}}$

Area $S = \pi r^2 \approx 3,142r^2, \quad S = \frac{\pi d^2}{4} \approx 0,785d^2, \quad S = \frac{Ud}{4}.$ (3.62)

Radius $r = \frac{U}{2\pi} \approx 0.159U.$ (3.63) **Diameter** $d = 2r = 2\sqrt{\frac{S}{\pi}} \approx 1.128\sqrt{S}.$ (3.64)

For the following formulas with angles see the definition of the angle in 3.1.1.2, p. 129.

Angle of Circumference (Fig. 3.25) $\alpha = \frac{1}{2}\overset{\frown}{BC} = \frac{1}{2}\sphericalangle B0C = \frac{1}{2}\varphi.$ (3.65a)

A Special Case is the Theorem of Thales (see p. 142) $\varphi = 180°$, i.e, $\alpha = 90°$. (3.65b)

Angle Between a Chord and a Tangent (Fig. 3.25) $\beta = \frac{1}{2}\overset{\frown}{AC} = \frac{1}{2}\sphericalangle C0A.$ (3.66)

Interior Angle (Fig. 3.26) $\gamma = \frac{1}{2}(\overset{\frown}{CB} + \overset{\frown}{ED}) = \frac{1}{2}(\sphericalangle B0C + \sphericalangle E0D).$ (3.67)

Exterior Angle (Fig. 3.27) $\alpha = \frac{1}{2}(\overset{\frown}{DE} - \overset{\frown}{BC}) = \frac{1}{2}(\sphericalangle E0C - \sphericalangle C0B).$ (3.68)

Angle Between Secant and Tangent (Fig. 3.27) $\beta = \frac{1}{2}(\overset{\frown}{TE} - \overset{\frown}{TB}) = \frac{1}{2}(\sphericalangle T0E - \sphericalangle B0T).$ (3.69)

Inscribed Angle (Fig. 3.28), D and E are arbitrary points on the arcs to the left and to the right.

$$\begin{aligned}\alpha &= \frac{1}{2}(B\widehat{D}C - C\widehat{E}B) = \frac{1}{2}(\sphericalangle B0C - \sphericalangle C0B)\\ &= \frac{1}{2}(360^\circ - \sphericalangle C0B - \sphericalangle C0B) = 180^\circ - \sphericalangle C0B\,.\end{aligned} \tag{3.70}$$

3.1.6.2 Circular Segment and Circular Sector

Defining quantities: Radius r and central angle α **(Fig. 3.29)**. The amounts to determine are:

Chord $$a = 2\sqrt{2hr - h^2} = 2r\sin\frac{\alpha}{2}\,. \tag{3.71}$$

Central Angle $$\alpha = 2\arcsin\frac{a}{2r} \quad (\alpha \text{ measured in degrees}). \tag{3.72}$$

Height of the Circular Segment $$h = r - \sqrt{r^2 - \frac{a^2}{4}} = r\left(1 - \cos\frac{\alpha}{2}\right) = \frac{a}{2}\tan\frac{\alpha}{4}\,. \tag{3.73}$$

Arc Length $$l = \frac{2\pi r\alpha}{360} \approx 0.01745 r\alpha \quad (\alpha \text{ in radian measure}) \tag{3.74a}$$

$$l \approx \frac{8b - a}{3} \quad \text{or} \quad l \approx \sqrt{a^2 + \frac{16}{3}h^2}\,. \tag{3.74b}$$

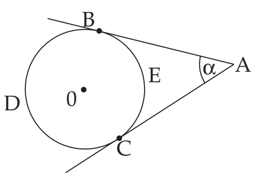


Figure 3.28

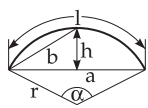


Figure 3.29

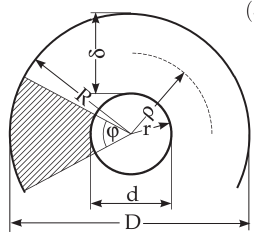


Figure 3.30

Area of the Sector $$S = \frac{\pi r^2\alpha}{360} \approx 0.00873 r^2\alpha. \tag{3.75}$$

Area of Circular Segment $$S = \frac{r^2}{2}\left(\frac{\pi\alpha}{180} - \sin\alpha\right) = \frac{1}{2}[lr - a(r-h)], \quad S \approx \frac{h}{15}(6a + 8b). \tag{3.76}$$

3.1.6.3 Annulus

Defining quantities of an annulus: Exterior radius R, interior radius r and central angle φ **(Fig. 3.30)**.

Exterior Diameter $D = 2\,R.$ (3.77) **Interior Diameter** $d = 2\,r.$ (3.78)

Mean Radius $\rho = \dfrac{R + r}{2}\,.$ (3.79) **Breadth of the Annulus** $\delta = R - r.$ (3.80)

Area of the Annulus $$S = \pi(R^2 - r^2) = \frac{\pi}{4}(D^2 - d^2) = 2\pi\,\rho\,\delta. \tag{3.81}$$

Area of an Annulus Sector for a Central Angle φ (shaded area in **Fig. 3.30**)

$$S_\varphi = \frac{\varphi\pi}{360}\left(R^2 - r^2\right) = \frac{\varphi\pi}{1440}\left(D^2 - d^2\right) = \frac{\varphi\,\pi}{180}\rho\,\delta \quad (\varphi \text{ in radian measure}). \tag{3.82}$$

3.2 Plane Trigonometry

3.2.1 Triangles

3.2.1.1 Calculations in Right-Angled Triangles in the Plane

1. Basic Formulas

Notation **(Fig. 3.31)**:
c hypotenuse; a, b other sides, or legs of the right angle; α and β the angles opposite to the sides a and b respectively; h altitude; p, q hypotenuse segments; S area.

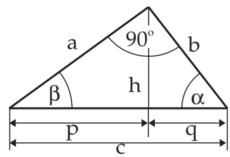


Figure 3.31

Sum of Angles $\alpha + \beta + \gamma = 180^0$ with $\gamma = 90°$, (3.83)

Calculation of Sides $a = c \sin\alpha = c\cos\beta = b\tan\alpha = b\cot\beta$, (3.84)

Pythagoras Theorem $a^2 + b^2 = c^2$. (3.85)

Thales Theorem The vertex angles of all triangles in a semicircle with hypotenuse as the base are right angles, i.e., all angles at circumference in this semicircle are right angles (see **Fig. 3.32** and (3.65b), p. 140).

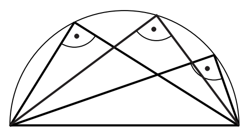


Figure 3.32

Euclidean Theorems $h^2 = p\,q,\quad a^2 = p\,c,\quad b^2 = q\,c,$ (3.86)

Area $S = \frac{ab}{2} = \frac{a^2}{2}\tan\beta = \frac{c^2}{4}\sin 2\beta.$ (3.87)

2. Calculation of Sides and Angles of a Right-Angled Triangle in the Plane

In a right-angled triangle among the six defining quantities (three angles α, β, γ and the sides opposite to them a, b, c, which are not all independent, of course), one angle, in **Fig. 3.31** the angle γ, is given as 90°.
A plane triangle can be determined by three defining quantities but these cannot be given arbitrarily (see 3.1.3.1, p. 133). So, in the case of a right-angled triangle only two more quantities can be given. The remaining three quantities can be determined from **Table 3.3** and (3.15) and (3.83).

Table 3.3 Defining quantities of a right-angled triangle in the plane

Given	**Calculation of the other quantities**		
e.g. a, α	$\beta = 90° - \alpha$	$b = a\cot\alpha$	$c = \frac{a}{\sin\alpha}$
e.g. b, α	$\beta = 90° - \alpha$	$a = b\tan\alpha$	$c = \frac{b}{\cos\alpha}$
e.g. c, α	$\beta = 90° - \alpha$	$a = c\sin\alpha$	$b = c\cos\alpha$
e.g. a, b	$\frac{a}{b} = \tan\alpha$	$c = \frac{a}{\sin\alpha}$	$\beta = 90° - \alpha$

3.2.1.2 Calculations in General (Oblique) Triangles in the Plane

1. Basic Formulas

Notation **(Fig. 3.33)**: a, b, c sides; α, β, γ the angles opposite to them; S area; R radius of the circumcircle; r radius of the incircle; $s = \frac{a+b+c}{2}$ half of the perimeter.

Cyclic Permutation

Because an oblique triangle has no distinguishing side or angle, from every formula containing the

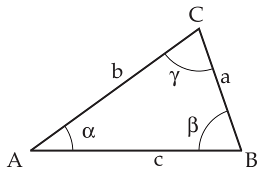


Figure 3.33

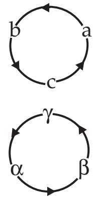


Figure 3.34

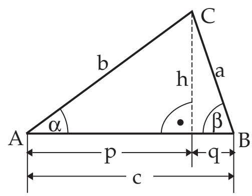


Figure 3.35

sides and angles it is possible to get two further formulas by cyclic permutation the sides and angles according to **Fig. 3.34**.

■ From $\frac{a}{b} = \frac{\sin\alpha}{\sin\beta}$ (sine law) one gets by cyclic permutation: $\frac{b}{c} = \frac{\sin\beta}{\sin\gamma}$, $\frac{c}{a} = \frac{\sin\gamma}{\sin\alpha}$.

Sine Law $$\frac{a}{\sin\alpha} = \frac{b}{\sin\beta} = \frac{c}{\sin\gamma} = 2R\,. \tag{3.88}$$

Projection Rule (see **Fig. 3.35**) $$c = a\cos\beta + b\cos\alpha\,. \tag{3.89}$$

Cosine Law or Pythagoras Theorem in General Triangles $$c^2 = a^2 + b^2 - 2ab\cos\gamma\,. \tag{3.90}$$

Mollweide Equations

$$(a+b)\sin\frac{\gamma}{2} = c\cos\left(\frac{\alpha-\beta}{2}\right), \tag{3.91a}$$

$$(a-b)\cos\frac{\gamma}{2} = c\sin\left(\frac{\alpha-\beta}{2}\right). \tag{3.91b}$$

Tangent Law $$\frac{a+b}{a-b} = \frac{\tan\frac{\alpha+\beta}{2}}{\tan\frac{\alpha-\beta}{2}}\,. \tag{3.92}$$

Half-Angle Formulas $$\tan\frac{\alpha}{2} = \sqrt{\frac{(s-b)\,(s-c)}{s\,(s-a)}}\,. \tag{3.93}$$

Tangent Formula $$\tan\alpha = \frac{a\sin\beta}{c - a\cos\beta} = \frac{a\sin\gamma}{b - a\cos\gamma}\,. \tag{3.94}$$

Additional Relations

$$\sin\frac{\alpha}{2} = \sqrt{\frac{(s-b)(s-c)}{bc}}\,, \tag{3.95a}$$

$$\cos\frac{\alpha}{2} = \sqrt{\frac{s(s-a)}{bc}}\,. \tag{3.95b}$$

Height Corresponding to the Side a $$h_a = b\sin\gamma = c\sin\beta\,. \tag{3.96}$$

Median of the Side a $$m_a = \frac{1}{2}\sqrt{b^2 + c^2 + 2bc\cos\alpha}\,. \tag{3.97}$$

Bisector of the Angle α $$l_\alpha = \frac{2bc\cos\frac{\alpha}{2}}{b+c}\,. \tag{3.98}$$

Radius of the Circumcircle $$R = \frac{a}{2\sin\alpha} = \frac{b}{2\sin\beta} = \frac{c}{2\sin\gamma}\,. \tag{3.99}$$

Radius of the Incircle $r = \sqrt{\frac{(s-a)(s-b)(s-c)}{s}} = s \tan\frac{\alpha}{2}\tan\frac{\beta}{2}\tan\frac{\gamma}{2}$ (3.100)

$$= 4R\sin\frac{\alpha}{2}\sin\frac{\beta}{2}\sin\frac{\gamma}{2}\,. \tag{3.101}$$

Area $S = \frac{1}{2}ab\sin\gamma = 2R^2\sin\alpha\sin\beta\sin\gamma = r\,s = \sqrt{s(s-a)(s-b)(s-c)}\,.$ (3.102)

The formula $S = \sqrt{s(s-a)(s-b)(s-c)}$ is called *Heron's formula.*

2. Calculation of Sides, Angles, and Area in General Triangles

According to the congruence theorems (see 3.1.3.2, p. 134) a triangle is determined by three independent quantities, among which there must be at least one side.

From here follow the four so-called *basic problems.* If from the six defining quantities (three angles α, β, γ and the sides opposite to them a, b, c) three independent quantities are given, one can calculate the remaining three with the equations in **Table 3.4**, p. 145.
In contrast to spherical trigonometry (see the second basic problem, **Table 3.9**, p. 172) in a plane triangle there is no way to get any side only from the angles.

3.2.2 Geodesic Applications

3.2.2.1 Geodesic Coordinates

In geometry usually *right-handed coordinate systems* are used to determine points in plane or space **(Fig. 3.170)**. In contrast with this, in geodesy *left-handed coordinate systems* are in use.

1. Geodesic Rectangular Coordinates

In a plane left-handed rectangular coordinate system **(Fig. 3.37)** the x-axis of abscissae is shown upward, the y-axis of ordinates is shown to the right. A point P has coordinates y_P, x_P. The orientation of the x-axis follows from practical reasons. When measuring long distances, for which mostly the Soldner- or the Gauss-Krueger coordinate systems are in use (see 3.4.1.2, p. 162), the positive x-axis points to *grid North*, the y-axis oriented to the right points to East. The enumeration of the quadrants follows a clockwise direction in contrast with the usual practice in geometry **(Fig. 3.37, Fig. 3.38)**.

If besides the position of a point in the plane also its altitude is to be considered, one can use a three-dimensional left-handed rectangular coordinate system (y, x, z), where the z-axis points to the *zenith* **(Fig. 3.36)**.

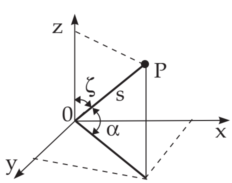


Figure 3.36

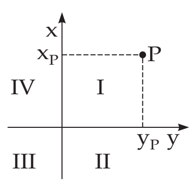


Figure 3.37

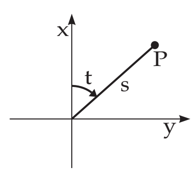


Figure 3.38

2. Geodesic Polar Coordinates

In the left-handed plane polar coordinate system of geodesy **(Fig. 3.38)** a point P is given by the *directional (azimuthal) angle* t between the axis of abscissae and the line segment s, and by the length of the line segment s between the point and the origin (called the pole). In geodesy the positive orientation of an angle is the clockwise direction.
To determine the altitude the *zenith distance* ζ can be used or the *vertical angle* respectively the *angle*

of tilt α. In **Fig. 3.36** is shown that in a three-dimensional rectangular left-handed coordinate system (see also left and right-handed coordinate systems 3.5.3.1, **2.**, p. 209), the zenith distance is measured between the zenith axis z and the line segment s, the angle of tilt between the line segment s and its perpendicular projection on the y, x plane.

Table 3.4 Defining quantities of a general triangle, basic problems

Given	Formulas for calculating the other quantities
1. 1 side and 2 angles (a, α, β)	$\gamma = 180° - \alpha - \beta, \quad b = \frac{a \sin\beta}{\sin\alpha},$ $c = \frac{a \sin\gamma}{\sin\alpha}, \quad S = \frac{1}{2}\, a\, b \sin\gamma\,.$
2. 2 sides and the angle between them (a, b, γ)	$\tan\frac{\alpha-\beta}{2} = \frac{a-b}{a+b}\cot\frac{\gamma}{2}, \quad \frac{\alpha+\beta}{2} = 90° - \frac{1}{2}\gamma\,;$ α and β come from $\alpha+\beta$ and $\alpha-\beta$, $c = \frac{a \sin\gamma}{\sin\alpha}, \quad S = \frac{1}{2}\, a\, b \sin\gamma\,.$
3. 2 sides and the angle opposite one of them (a, b, α)	$\sin\beta = \frac{b\sin\alpha}{a}.$ If $a \geq b$ holds, then $\beta < 90°$ and is uniquely determined. If $a < b$ holds, the following cases occur: 1. β has two values for $b\sin\alpha < a$ $(\beta_2 = 180° - \beta_1)$. 2. β has exactly one value $(90°)$ for $b\sin\alpha = a$. 3. For $b\sin\alpha > a$ there is no such triangle. $\gamma = 180° - (\alpha+\beta), \quad c = \frac{a\sin\gamma}{\sin\alpha}, \quad S = \frac{1}{2}\, a\, b\sin\gamma.$
4. 3 sides (a, b, c)	$r = \sqrt{\frac{(s-a)(s-b)(s-c)}{s}},$ $\tan\frac{\alpha}{2} = \frac{r}{s-a}, \quad \tan\frac{\beta}{2} = \frac{r}{s-b}, \quad \tan\frac{\gamma}{2} = \frac{r}{s-c},$ $S = r\,s = \sqrt{s\,(s-a)(s-b)(s-c)}\,.$

3. Scale

In cartography the scale factor M is the ratio of a segment s_{K1} in a coordinate system K_1 with respect to the corresponding segment s_{K2} in another coordinate system K_2.

1. Conversion of Segments With m as a *modulus* or *scale* and N as an index for the nature and K as the index of the map holds:

$$M = 1 : m = s_K : s_N\,. \tag{3.103a}$$

For two segments s_{K1}, s_{K2} with different moduli m_1, m_2 yields:

$$s_{K1} : s_{K2} = m_2 : m_1. \tag{3.103b}$$

2. Conversion of Areas If the areas are calculated according to the formulas $F_K = a_K b_K,\ F_N = a_N b_N$, then:

$$F_N = F_K m^2\,. \tag{3.104a}$$

For two areas F_1, F_2 with different moduli m_1, m_2 :

$$F_{K1} : F_{K2} = m_2^2 : m_1^2. \tag{3.104b}$$

3.2.2.2 Angles in Geodesy

1. Grade or Gon Division

In geodesy, in contrast to mathematics (see 3.1.1.5, p. 131) the *gon* measure is in use as a unit for angles. The perigon or full angle corresponds here to 400 grade or gon. The conversion between degrees and gons can be performed by the formulas in **Table 3.5**:

Table 3.5 Grade and Gon Division

1 Full angle	$= 360°$	$= 2\pi$ rad	$= 400$ gon
1 right angle	$= 90°$	$= \frac{\pi}{2}$ rad	$= 100$ gon
1 gon		$= \frac{\pi}{200}$ rad	$= 1000$ mgon

2. Directional Angle

The *directional angle* t at a point P gives the direction of an oriented line segment with respect to a line passing through the point P parallel to the x-axis (see point A and the directional angle t_{AB} in **Fig. 3.39**). Because the measuring of angles in geodesy is made in a clockwise direction **(Fig. 3.37, Fig. 3.38)**, the quadrants are enumerated in the opposite order to the right-handed Cartesian coordinate system of plane trigonometry (**Table 3.6**). The formulas of plane trigonometry are valid without change.

Table 3.6 Directional angle in a segment with correct sign for arctan

Quadrant	**I**	**II**	**III**	**IV**
Sign of numerator	+	−	−	+
$\frac{\Delta y}{\Delta x}$	$\tan > 0$	$\tan < 0$	$\tan > 0$	$\tan < 0$
Directional angle t	t_0 gon	$t_0 + 200$ gon	$t_0 + 200$ gon	$t_0 + 400$ gon

3. Coordinate Transformations

1. Calculation of Polar Coordinates from Rectangular Coordinates For two points $A(y_A, x_A)$ and $B(y_B, x_B)$ in a rectangular coordinate system **(Fig. 3.39)** with the segment s_{AB} oriented from A to B and the directional angles t_{AB}, t_{BA} the following formulas are valid:

$$\frac{y_B - y_A}{x_B - x_A} = \frac{\Delta y_{AB}}{\Delta x_{AB}}, \tag{3.105a}$$

$$s_{AB} = \sqrt{\Delta y_{AB}^2 + \Delta x_{AB}^2}, \tag{3.105b}$$

$$\tan t_{AB} = \frac{\Delta y_{AB}}{\Delta x_{AB}}, \tag{3.105c}$$

$$t_{BA} = t_{AB} \pm 200\,\text{gon}. \tag{3.105d}$$

The quadrant of the angle t depends on the signs of Δy_{AB} and Δx_{AB}. If using a calculator $\frac{\Delta y}{\Delta x}$ is punched in with the correct signs for Δy and Δx, then one gets an angle t_0 by pressing the arctan button to which is to add the gon-value given in **Table 3.6** according to the corresponding quadrant.

2. Calculation of Rectangular Coordinates from Distances and Angles In a rectangular coordinate system are to determine the coordinates of a point C by measuring in a local polar system **(Fig. 3.40)**.

Given: $y_A, x_A; y_B, x_B$. **Measured:** α, s_{BC}. **Find:** y_C, x_C.
Solution:

$$\tan t_{AB} = \frac{\Delta y_{AB}}{\Delta x_{AB}}, \quad (3.106a) \qquad t_{BC} = t_{AB} + \alpha \pm 200\,\text{gon}, \quad (3.106b)$$

$$y_C = y_B + s_{BC} \sin t_{BC}, \quad (3.106c) \qquad x_C = x_B + s_{BC} \cos t_{BC}. \quad (3.106d)$$

If also s_{AB} is measured, then the difference between the locally measured distance and the distance computed from the coordinates can be considered by multiplying with the *scale factor* q, where q must be very close to 1:

$$q = \frac{\text{calculated distance}}{\text{measured distance}} = \frac{\sqrt{\Delta y_{AB}^2 + \Delta x_{AB}^2}}{s_{AB}}, \quad (3.107a)$$

$$y_C = y_B + s_{BC} q \sin t_{BC}, \quad (3.107b) \qquad x_C = x_B + s_{BC} q \cos t_{BC}. \quad (3.107c)$$

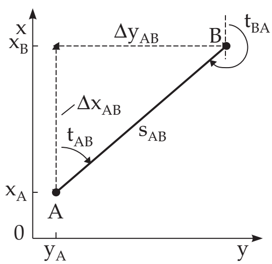


Figure 3.39

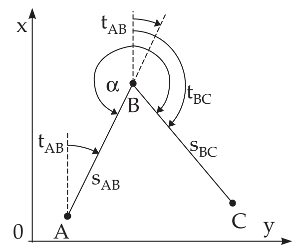


Figure 3.40

3. Coordinate Transformation Between Two Rectangular Coordinate Systems In order to locate a given point on a country-map the local system y', x' is to be transformed into the system y, x of coordinates of the map **(Fig. 3.41)**. The system y', x' is rotated into y, x by an angle φ and is translated parallel by y_0, x_0. The directional angles in the system y', x' are denoted by ϑ. The coordinates of A and B are given in both systems and the coordinates of a point C in the x', y'–system. The transformation is given by the following relations:

$$s_{AB} = \sqrt{\Delta y_{AB}^2 + \Delta x_{AB}^2}, \quad (3.108a) \qquad s'_{AB} = \sqrt{\Delta y'^2_{AB} + \Delta x'^2_{AB}}, \quad (3.108b)$$

$$q = \frac{s_{AB}}{s'_{AB}}, \quad (3.108c) \qquad \varphi = t_{AB} - \vartheta_{AB}, \quad (3.108d)$$

$$\tan t_{AB} = \frac{\Delta y_{AB}}{\Delta x_{AB}}, \quad (3.108e) \qquad \tan \vartheta_{AB} = \frac{\Delta y'_{AB}}{\Delta x'_{AB}}, \quad (3.108f)$$

$$y_0 = y_A - q x_A \sin\varphi - q y_A \cos\varphi, \quad (3.108g) \qquad x_0 = x_A + q y_A \sin\varphi - q x_A \cos\varphi, \quad (3.108h)$$

$$y_C = y_A + q \sin\varphi (x'_C - x'_A) + q \cos\varphi (y'_C - y'_A), \quad (3.108i)$$
$$x_C = x_A + q \cos\varphi (x'_C - x'_A) - q \sin\varphi (y'_C - y'_A). \quad (3.108j)$$

Remark: The following two formulas can be used as a check.

$$y_C = y_A + q s'_{AC} \sin(\varphi + \vartheta_{AC}), \quad (3.108k) \qquad x_C = x_A + q s'_{AC} \cos(\varphi + \vartheta_{AC}). \quad (3.108l)$$

If the segment AB is on the x'-axis, the formulas reduce to

$$a = \frac{\Delta y_{AB}}{y'_B} = q\sin\varphi, \quad (3.109a) \qquad b = \frac{\Delta x_{AB}}{x'_B} = q\cos\varphi, \quad (3.109b)$$

$$y_C = y_A + ax'_C + by'_C, \quad (3.109c) \qquad x_C = x_A + bx'_C - ay'_C, \quad (3.109d)$$

$$y'_C = \Delta y_{AC} b - \Delta x_{AC} a, \quad (3.109e) \qquad x'_C = \Delta x_{AC} b + \Delta y_{AC} a. \quad (3.109f)$$

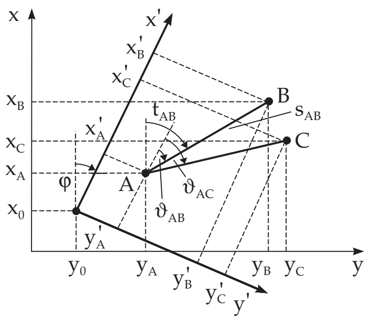


Figure 3.41

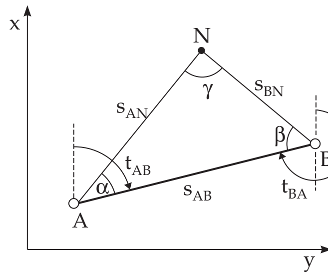


Figure 3.42

3.2.2.3 Applications in Surveying

The determination of the coordinates of a point N to be fixed by *triangulation* is a frequent measuring problem in geodesy. The methods of solving it are *intersection*, *three-point resection*, *arc intersection*, *free stationing* and *traversing*. The last two methods are not discussed here.

1. Intersection

1. Intersection of Two Oriented Lines or *first fundamental problem of triangulation*: To determine a point N by two given points A and B with the help of a triangle ABN **(Fig. 3.42)**.

Given: $y_A, x_A; y_B, x_B$. **Measured:** α, β, if it is possible also γ or $\gamma = 200$ gon $-\alpha - \beta$. **Find:** y_N, x_N.
Solution:

$$\tan t_{AB} = \frac{\Delta y_{AB}}{\Delta x_{AB}}, \quad (3.110a) \qquad s_{AB} = \sqrt{\Delta y_{AB}^2 + \Delta x_{AB}^2} = |\Delta y_{AB}\sin t_{AB}| + |\Delta x_{AB}\cos t_{AB}|, \quad (3.110b)$$

$$s_{BN} = s_{AB}\frac{\sin\alpha}{\sin\gamma} = s_{AB}\frac{\sin\alpha}{\sin(\alpha+\beta)}, \quad (3.110c) \qquad s_{AN} = s_{AB}\frac{\sin\beta}{\sin\gamma} = s_{AB}\frac{\sin\beta}{\sin(\alpha+\beta)}, \quad (3.110d)$$

$$t_{AN} = t_{AB} - \alpha, \quad (3.110e) \qquad t_{BN} = t_{BA} + \beta = t_{AB} + \beta \pm 200\,\text{gon}, \quad (3.110f)$$

$$y_N = y_A + s_{AN}\sin t_{AN} = y_B + s_{BN}\sin t_{BN}, \quad (3.110g)$$

$$y_N = x_A + s_{AN}\cos t_{AN} = x_B + s_{BN}\cos t_{BN}. \quad (3.110h)$$

2. Intersection Problem for Non-Visible *B* If the point B cannot be seen from A, the directional angles t_{AN} and t_{BN} are to be determined with respect to reference directions to other visible points D and E whose coordinates are known **(Fig. 3.43)**.
Given: $y_A, x_A; y_B, x_B; y_D, x_D; y_E, x_E$. **Measured:** δ in A, ε in B, and if it is possible, also γ.
Find: y_N, x_N.

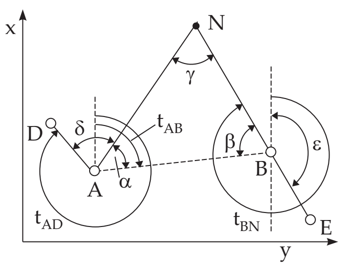


Figure 3.43

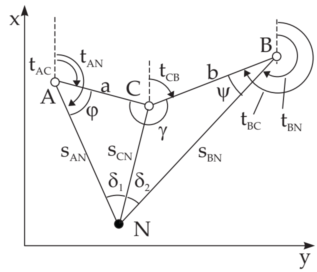


Figure 3.44

Solution: Reducing to the first fundamental problem, calculating $\tan t_{AB}$, according to (3.110b) yields:

$$\tan t_{AD} = \frac{\Delta y_{AD}}{\Delta x_{AD}}, \qquad (3.111a) \qquad\qquad \tan t_{BE} = \frac{\Delta y_{EB}}{\Delta x_{EB}}, \qquad (3.111b)$$

$$t_{AN} = t_{AD} + \delta, \qquad (3.111c) \qquad\qquad t_{BN} = t_{BE} + \varepsilon, \qquad (3.111d)$$

$$\alpha = t_{AB} - t_{AN}, \qquad (3.111e) \qquad\qquad \beta = t_{BN} - t_{BA}, \qquad (3.111f)$$

$$\tan t_{AN} = \frac{\Delta y_{NA}}{\Delta x_{NA}}, \qquad (3.111g) \qquad\qquad \tan t_{BN} = \frac{\Delta y_{NB}}{\Delta x_{NB}}, \qquad (3.111h)$$

$$x_N = \frac{\Delta y_{BA} + x_A \tan t_{AN} - x_B \tan t_{BN}}{\tan t_{AN} - \tan t_{BN}}, \quad (3.111i) \qquad y_N = y_B + (x_N - x_B)\tan t_{BN}. \quad (3.111j)$$

2. Three-Point Resection

1. Snellius Problem of Three-Point Resection or to determine a point N by three given points A, B, C; also called the *second fundamental problem of triangulation* **(Fig. 3.44)**:

Given: $y_A, x_A; y_B, x_B; y_C, x_C$. **Measured:** δ_1, δ_2 in N. **Find:** y_N, x_N.

Solution:

$$\tan t_{AC} = \frac{\Delta y_{AC}}{\Delta x_{AC}}, \qquad (3.112a) \qquad\qquad \tan t_{BC} = \frac{\Delta y_{BC}}{\Delta x_{BC}}, \qquad (3.112b)$$

$$a = \frac{\Delta y_{AC}}{\sin t_{AC}} = \frac{\Delta x_{AC}}{\cos t_{AC}}, \qquad (3.112c) \qquad\qquad b = \frac{\Delta y_{BC}}{\sin t_{BC}} = \frac{\Delta x_{BC}}{\cos t_{BC}}, \qquad (3.112d)$$

$$\gamma = t_{CA} - t_{CB} = t_{AC} - t_{BC}, \qquad (3.112e) \qquad\qquad \frac{\varphi + \psi}{2} = 180^0 - \frac{\gamma + \delta_1 + \delta_2}{2}, \qquad (3.112f)$$

The equality (3.112f) is a first condition to determine φ and ψ. A second condition one gets from (2.114) and (2.115), p. 83:

$$\frac{\sin\varphi + \sin\psi}{\sin\varphi - \sin\psi} = \tan\frac{\varphi + \psi}{2} \cdot \cot\frac{\varphi - \psi}{2}. \qquad (3.112g)$$

With the sine law (3.88) follows

$$\frac{\sin\varphi}{\sin\delta_1} = \frac{s_{CN}}{a}, \quad \frac{\sin\psi}{\sin\delta_2} = \frac{s_{CN}}{b}, \tag{3.112h}$$

putting into (3.112g) gives

$$\tan\frac{\varphi-\psi}{2} = \tan\frac{\varphi+\psi}{2}\cdot\frac{\sin\varphi-\sin\psi}{\sin\varphi+\sin\psi} = \tan\frac{\varphi+\psi}{2}\cdot\frac{b\sin\delta_1 - a\sin\delta_2}{b\sin\delta_1 + a\sin\delta_2}. \tag{3.112i}$$

From (3.112i) one gets $\frac{\varphi-\psi}{2}$ and together with (3.112f)

$$\varphi = \frac{\varphi+\psi}{2} + \frac{\varphi-\psi}{2}, \quad \psi = \frac{\varphi+\psi}{2} - \frac{\varphi-\psi}{2}. \tag{3.112j}$$

From here one can determine the following line segments and points:

$$s_{AN} = \frac{a}{\sin\delta_1}\sin(\delta_1+\varphi), \quad (3.112k) \qquad s_{BN} = \frac{b}{\sin\delta_2}\sin(\delta_2+\psi), \tag{3.112l}$$

$$s_{CN} = \frac{a}{\sin\delta_1}\sin\varphi = \frac{b}{\sin\delta_2}\sin\psi, \tag{3.112m}$$

$$x_N = x_A + s_{AN}\cos t_{AN} = x_B + s_{BN}\cos t_{BN}, \tag{3.112n}$$

$$y_N = y_A + s_{AN}\sin t_{AN} = y_B + s_{BN}\sin t_{BN}. \tag{3.112o}$$

2. Three-Point Resection by Cassini

Given: $y_A, x_A; y_B, x_B; y_C, x_C$. **Measured:** δ_1, δ_2 in N. **Find:** y_N, x_N.

For this method two reference points P and Q are to be used, which are on the reference circles passing through A, C, P and B, C, Q so that both are on a line containing N **(Fig. 3.45)**. The centers of the circles H_1 and H_2 are at the intersection points of the mid perpendicular of $\overline{AC}$ and of $\overline{BC}$ with the segments PC and QC. The angles δ_1, δ_2 measured at N appear also at P and Q as angles of circumference.

Solution:

$$y_P = y_A + (x_C - x_A)\cot\delta_1, \quad (3.113a) \qquad x_P = x_A + (y_C - y_A)\cot\delta_1, \tag{3.113b}$$

$$y_Q = y_B + (x_B - x_C)\cot\delta_2, \quad (3.113c) \qquad x_Q = x_B + (y_B - y_C)\cot\delta_2, \tag{3.113d}$$

$$\cot t_{PQ} = \frac{\Delta y_{PQ}}{\Delta x_{PQ}}, \quad (3.113e) \qquad x_N = x_P + \frac{y_C - y_P + (x_C - x_P)\cot t_{PQ}}{\tan t_{PQ} + \cot t_{PQ}}, \tag{3.113f}$$

$$y_N = y_P + (x_N - x_P)\tan t_{PQ} \qquad (\tan t_{PQ} < \cot t_{PQ}), \tag{3.113g}$$

$$y_N = y_C - (x_N - x_C)\cot t_{PQ} \qquad (\cot t_{PQ} < \tan t_{PQ}), \tag{3.113h}$$

Dangerous Circle: When choosing the points it is to be ensured that they do not lie on one circle, because then there is no solution; one talks about a so-called *dangerous circle*. The closer the points are near to the dangerous circle, the less gets the accuracy of the method.

3. Arc Intersection

With this method a so-called new point N is to be determined as intersection point of two arcs around two points A and B with known coordinates and with measured radii s_{AN} and s_{BN} **(Fig. 3.46)**. The unknown line segment s_{AB} is calculated and the angles are to be calculated from the now already known three sides of the triangle ABN.

A second solution – not discussed here – starts from the decomposition of the general triangle into two

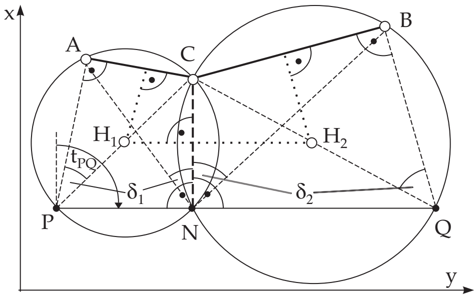


Figure 3.45

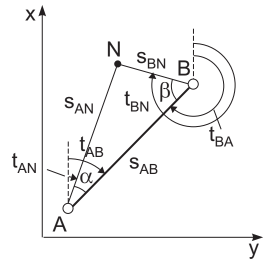


Figure 3.46

right-angled triangles.

Given: $y_A, x_A : y_B, x_B$. **Measured:** $s_{AN}; s_{BN}$. **Find:** s_{AB}, y_N, x_N.
Solution:

$$s_{AB} = \sqrt{\Delta y_{AB}^2 + \Delta x_{AB}^2}\,, \qquad (3.114a)$$

$$\tan t_{AB} = \frac{\Delta y_{AB}}{\Delta x_{AB}}\,, \qquad (3.114b)$$

$$t_{BA} = t_{AB} + 200\,\text{gon}, \qquad (3.114c)$$

$$\cos\alpha = \frac{s_{AN}^2 + s_{AB}^2 - s_{BN}^2}{2 s_{AN} s_{AB}}\,, \qquad (3.114d)$$

$$\cos\beta = \frac{s_{BN}^2 + s_{AB}^2 - s_{AN}^2}{2 s_{BN} s_{AB}}\,, \qquad (3.114e)$$

$$t_{AN} = t_{AB} - \alpha, \qquad (3.114f)$$

$$t_{BN} = t_{BA} - \beta, \qquad (3.114g)$$

$$y_N = y_A + s_{AN}\sin t_{AN}, \qquad (3.114h)$$

$$x_N = x_A + s_{AN}\cos t_{AN}, \qquad (3.114i)$$

$$y_N = y_B + s_{BN}\sin t_{BN}, \qquad (3.114j)$$

$$x_N = x_B + s_{BN}\cos t_{BN}. \qquad (3.114k)$$

3.3 Stereometry

3.3.1 Lines and Planes in Space

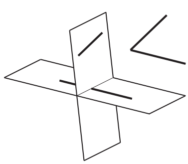


Figure 3.47

1. Two Lines Two lines in the same plane have either one or no common point. In the second case they are *parallel.* If there is no plane such that it contains both lines, they are *skew lines.* The *angle between two skew lines* is defined by the angle between two lines parallel to them and passing through a common point **(Fig. 3.47)**. The distance between two skew lines is defined by the segment which is perpendicular to both of them. (There is always a unique transversal line which is perpendicular to both skew lines and intersects them, too.)

2. Two Planes Two planes intersect each other in a line or they have no common point. In this second case they are parallel. If two planes are perpendicular to the same line or if both are parallel to every intersecting pair of lines in the other, the planes are parallel.

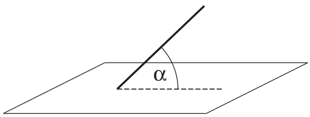


Figure 3.48

3. Line and Plane A line can be incident to a plane, it can have one or no common point with the plane. In this last case it is parallel to the plane. The angle between a line and a plane is measured by the angle between the line and its orthogonal projection on the plane **(Fig. 3.48)**. If this projection is only a point, i.e., if the line is perpendicular to two different intersecting lines in the plane, then the line is *perpendicular* or *orthogonal* to the plane.

3.3.2 Edge, Corner, Solid Angle

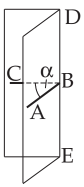


Figure 3.49

1. Edge or *dihedral angle* is a figure formed by two infinite half-planes starting at the same line **(Fig.3.49)**. In everyday terminology the word edge is used for the intersection line of the two half-planes. As a measure of edges, one uses the *edge-angle* ABC, the angle between two half-lines lying in the half-planes and being perpendicular to the intersection line DE at a point B.

2. Corner or *polyhedral angle* $0ABCDE$ **(Fig. 3.50)** is a figure formed by several planes, the *lateral faces*, which go through a common point, the vertex 0, and intersect each other at the lines $0A$, $0B$, ...

Two lines which bound the same lateral face form a *plane angle*, while neighboring faces form a *dihedral angle*.

Polyhedra are equal to each other, i.e., they are *congruent*, if they are superposable. For this the corresponding elements, i.e., the edges and plane angles at the vertex must be coincident. If the corresponding elements at the vertex are equal, but they have an opposite order of sequence, the corners are not superposable, and they are called *symmetric corners*, because they can be brought into a symmetric position as shown in **Fig. 3.51**.

A *convex polyhedral angle* lies completely on one side of each of its faces.

The sum of the plane angles $A0B + B0C + \ldots + E0A$ **(Fig. 3.50)** is less than 360° for every convex polyhedron.

3. Trihedral Angles are congruent if they coincide in the following elements:

- in two faces and the corresponding dihedral angle,
- in one face and both dihedral angles belonging to it,
- in three corresponding faces in the same order of sequence,
- in three corresponding dihedral angles in the same order of sequence.

4. Solid Angle A pencil of rays starting from the same point (and intersecting a closed curve) forms a solid angle in space **(Fig. 3.52)**. It is denoted by Ω and calculated by the equality

$$\Omega = \frac{S}{r^2}\,. \tag{3.115a}$$

Here S means the piece of the spherical surface cut out by the solid angle from a sphere whose radius is r, and whose center is at the vertex of the solid angle. The unit of solid angle is the *steradian* (sr) (see also p. 1055):

$$1\,\mathrm{sr} = \frac{1\,\mathrm{m}^2}{1\,\mathrm{m}^2}\,, \tag{3.115b}$$

i.e., a solid angle of 1 sr cuts out a surface area of $1\,\mathrm{m}^2$ of the unit sphere ($r = 1\,\mathrm{m}$).

■ **A:** The full solid angle is $\Omega = 4\pi r^2/r^2 = 4\pi$.

■ **B:** A cone with a vertex angle (also called an apex angle) $\alpha = 120°$ defines(determines) a solid angle $\Omega = 2\pi r^2(1 - \cos(\alpha/2))/r^2 = \pi$, where the formula for a spherical cap (3.163)has been used.

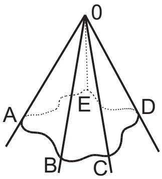


Figure 3.50

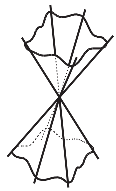


Figure 3.51

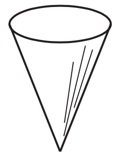


Figure 3.52

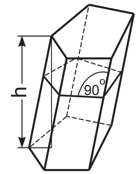


Figure 3.53

3.3.3 Polyeder or Polyhedron

In this paragraph the following notations are used : V volume, S total surface, M lateral area, h altitude, A_G base area.

1. Polyhedron is a solid bounded by plane polygons.

2. Prism (Fig. 3.53) is a polyhedron with two congruent bases, and having parallelograms as additional faces. A *right prism* has edges perpendicular to the base, a *regular prism* is a right prism with a regular polygon as the base. For the prism

$$V = A_G h, \quad (3.116) \qquad M = p\,l, \quad (3.117) \qquad S = M + 2A_G. \quad (3.118)$$

holds. Here p is the perimeter of the section perpendicular to the edges, and l is the length of the edges. If the edges are still parallel to each other but the bases are not, the lateral faces are trapezoids. If the bases of a triangular prism are not parallel to each other, its volume can be calculated by the formula **(Fig. 3.54)**:

$$V = \frac{(a+b+c)Q}{3}\,, \qquad (3.119)$$

where Q is a perpendicular cut, a, b, and c are the lengths of the parallel edges. If the bases of the prism are not parallel, then its volume is

$$V = l\,Q, \qquad (3.120)$$

where l is the length of the line segment $\overline{BC}$ connecting the centers of gravity of the bases, and Q is the cross-cut perpendicular to this line.

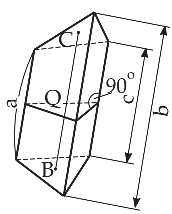


Figure 3.54

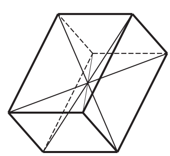


Figure 3.55

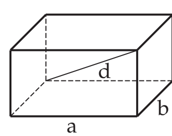


Figure 3.56

3. Parallelepiped is a prism with parallelograms as bases **(Fig. 3.55)**, i.e., it is bounded by six parallelograms. In a parallelepiped all the four body diagonals intersect each other at the same point, the midpoint, and halve each other.

4. Rectangular Parallelepiped or *block* is a right parallelepiped with rectangles as bases. In a block **(Fig. 3.56)** the body diagonals have the same length. If a, b, and c are the edge lengths of the block and d is the length of the diagonal, then

$$d^2 = a^2 + b^2 + c^2, \quad (3.121) \qquad V = a\,b\,c, \quad (3.122) \qquad S = 2(a\,b + b\,c + c\,a). \quad (3.123)$$

5. Cube (or *regular hexahedron*) is a block with equal edge lengths: $a = b = c$,

$$d^2 = 3\,a^2, \quad (3.124) \qquad V = a^3, \quad (3.125) \qquad S = 6\,a^2. \quad (3.126)$$

6. Pyramid (Fig. 3.57) is a polyhedron whose base is a polygon and its lateral faces are triangles with a common point, the vertex. A pyramid is called *right* if the foot of the perpendicular from the vertex to the base A_G is at the midpoint of the base. It is called *regular* if it is right and the base is a regular polygon **(Fig. 3.58)**, and n *faced* if the base is an n gon. Together with the base the pyramid has $(n+1)$ faces. For the volume

$$V = \frac{A_G\,h}{3} \quad (3.127)$$

holds. For the lateral area of the regular pyramid

$$M = \frac{1}{2} p h_s \quad (3.128)$$

holds with p as the perimeter of the base and h_s as the altitude of a face.

7. Frustum of a Pyramid or *truncated pyramid* is a pyramid whose vertex is cut away by a plane parallel to the base **(Fig. 3.57, Fig. 3.59)**. Denoting by $\overline{S0}$ the altitude of the pyramid, i.e., the perpendicular from the vertex to the base, then

$$\frac{\overline{SA_1}}{\overline{A_1A}} = \frac{\overline{SB_1}}{\overline{B_1B}} = \frac{\overline{SC_1}}{\overline{C_1C}} = \ldots = \frac{\overline{S0_1}}{\overline{0_10}}, \quad (3.129)$$

$$\frac{\text{Area } ABCDEF}{\text{Area } A_1B_1C_1D_1E_1F_1} = \left(\frac{\overline{S0}}{\overline{S0_1}}\right)^2. \quad (3.130)$$

holds. If A_D and A_G are the upper and lower bases, resp., h is the altitude of the truncated pyramid, i.e., the distance between the bases, and a_D and a_G are corresponding sides of the bases, then

$$V = \frac{1}{3} h \left[A_G + A_D + \sqrt{A_G A_D} \right] = \frac{1}{3} h A_G \left[1 + \frac{a_D}{a_G} + \left(\frac{a_D}{a_G}\right)^2 \right]. \quad (3.131)$$

The lateral surface of a regular truncated pyramid is

$$M = \frac{p_D + p_G}{2} h_s\,, \quad (3.132)$$

where p_D and p_G are the perimeters of the bases, and h_s is the altitude of the faces.

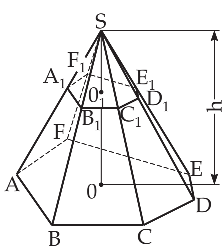


Figure 3.57

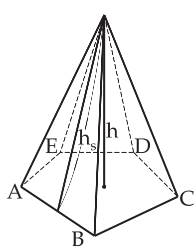


Figure 3.58

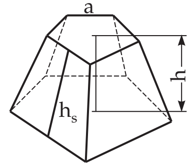


Figure 3.59

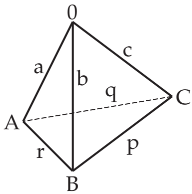


Figure 3.60

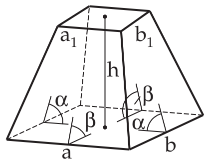


Figure 3.61

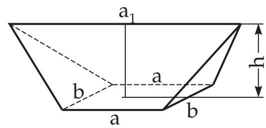


Figure 3.62

8. Tetrahedron is a triangular pyramid **(Fig. 3.60)**. With the notation $\overline{0A} = a,\ \overline{0B} = b,\ \overline{0C} = c,\ \overline{CA} = q,\ \overline{BC} = p$ and $\overline{AB} = r$ the following holds:

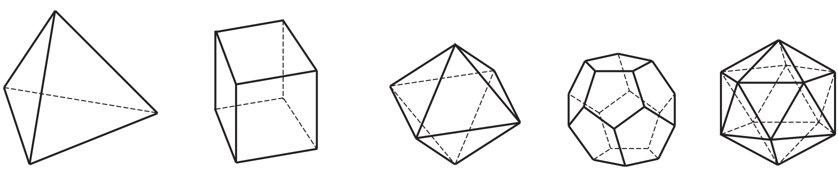
$$V^2 = \frac{1}{288} \begin{vmatrix} 0 & r^2 & q^2 & a^2 & 1 \\ r^2 & 0 & p^2 & b^2 & 1 \\ q^2 & p^2 & 0 & c^2 & 1 \\ a^2 & b^2 & c^2 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 \end{vmatrix} . \tag{3.133}$$

9. Obelisk is a polyhedron whose lateral faces are all trapezoids. In the special case in **Fig. 3.61** the bases are rectangles, the opposite edges have the same inclination to the base but they do not have a common point. If a, b and a_1, b_1 are the sides of the bases of the obelisk and h is the altitude of it, then:

$$V = \frac{h}{6}\left[(2a + a_1)\,b + (2a_1 + a)\,b_1\right] = \frac{h}{6}\left[ab + (a + a_1)\,(b + b_1) + a_1 b_1\right]. \tag{3.134}$$

10. Wedge is a polyhedron whose base is a rectangle, its lateral faces are two opposite isosceles triangles and two isosceles trapezoids **(Fig. 3.62)**. For the volume holds

$$V = \frac{1}{6}\,(2a + a_1)\,b\,h\,. \tag{3.135}$$



a) b) c) d) e)

Figure 3.63

11. Regular Polyeder have congruent regular polyeders as faces and congruent regular corners. The five possible regular polyheders are represented in **Fig. 3.63**; **Table 3.7** shows the corresponding data.

12. Euler's Theorem on Polyeders If e is the number of vertices, f is the number of faces, and k is the number of edges of a convex polyhedron then

$$e - k + f = 2\,. \tag{3.136}$$

Examples are given in **Table 3.7**.

Table 3.7 Regular polyeders with edge length a

Name	**Number and form of faces**	**Number of**		**Total area**	**Volume**
		Edges	**Corners**	S/a^2	V/a^3
Tetrahedron	4 triangles	6	4	$\sqrt{3} = 1.7321$	$\frac{\sqrt{2}}{12} = 0.1179$
Cube	6 squares	12	8	$6 = 6.0$	$1 = 1.0$
Octahedron	8 triangles	12	6	$2\sqrt{3} = 3.4641$	$\frac{\sqrt{2}}{3} = 0.4714$
Dodecahedron	12 pentagons	30	20	$3\sqrt{5(5+2\sqrt{5})} = 20.6457$	$\frac{15+7\sqrt{5}}{4} = 7.6631$
Icosahedron	20 triangles	30	12	$5\sqrt{3} = 8.6603$	$\frac{5(3+\sqrt{5})}{12} = 2.1817$

3.3.4 Solids Bounded by Curved Surfaces

In this paragraph the following notations are used: V volume, S total surface, M lateral surface, h altitude, A_G base.

1. Cylindrical Surface is a curved surface which can be got by parallel translation of a line, the *generating line* or *generator* along a curve, the so-called *directing curve* **(Fig. 3.64)**.

2. Cylinder is a solid bounded by a cylindrical surface with a closed directing curve, and by two parallel bases cut out from two parallel planes by the cylindrical surface. For every arbitrary cylinder **(Fig. 3.65)** with the base perimeter p, with the perimeter s of the cut perpendicular to the apothem, whose area is Q, and with the length l of the apothem the following is valid:

$$V = A_G\, h = Q\, l, \quad (3.137) \qquad M = p\, h = s\, l. \quad (3.138)$$

3. Right Circular Cylinder has a circle as base, and its apothems are perpendicular to the plane of the circle **(Fig. 3.66)**. With a base radius R

$$V = \pi R^2 h, \quad (3.139) \qquad M = 2\pi R h, \quad (3.140) \qquad S = 2\pi R(R+h). \quad (3.141)$$

4. Obliquely Truncated Cylinder (Fig. 3.67)

$$V = \pi R^2 \frac{h_1 + h_2}{2}, \quad (3.142) \qquad M = \pi R\,(h_1 + h_2)\,, \quad (3.143)$$

$$S = \pi R \left[h_1 + h_2 + R + \sqrt{R^2 + \left(\frac{h_2 - h_1}{2}\right)^2}\,\right]. \quad (3.144)$$

5. Ungula of the Cylinder With the notation of **Fig. 3.68** and with $\alpha = \varphi/2$ in radians

$$V = \frac{h}{3b}\left[a(3R^2 - a^2) + 3R^2(b-R)\alpha\right] = \frac{hR^3}{b}\left(\sin\alpha - \frac{\sin^3\alpha}{3} - \alpha\cos\alpha\right), \quad (3.145)$$

$$M = \frac{2Rh}{b}\,[(b-R)\alpha + a]\,, \quad (3.146)$$

where the formulas are valid even in the case $b > R$, $\varphi > \pi$.

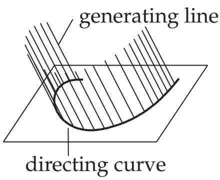


Figure 3.64

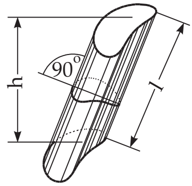


Figure 3.65

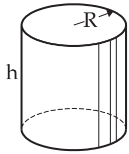


Figure 3.66

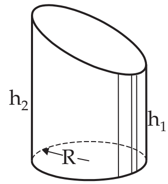


Figure 3.67

6. Hollow Cylinder With the notation R for the outside radius and r for the inside one, $\delta = R - r$ for the difference of the radii, and $\varrho = \dfrac{R+r}{2}$ for the mean radius **(Fig. 3.69)**

$$V = \pi\, h(R^2 - r^2) = \pi\, h\delta(2\,R - \delta) = \pi\, h\delta(2\,r + \delta) = 2\,\pi\, h\delta\varrho. \tag{3.147}$$

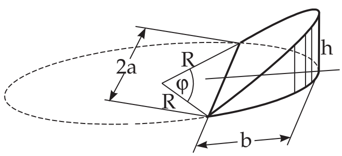


Figure 3.68

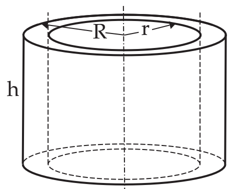


Figure 3.69

7. Conical Surface arises by moving a line, the generating line, along a curve, the direction curve so that it always goes through a fixed point, the vertex **(Fig. 3.70)**.

8. Cone (Fig. 3.71) is bounded by a conical surface with a closed direction curve and a base cut out from a plane by the surface. For an arbitrary cone

$$V = \frac{h\,A_G}{3}. \tag{3.148}$$

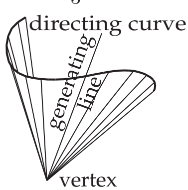


Figure 3.70

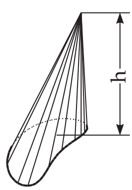


Figure 3.71

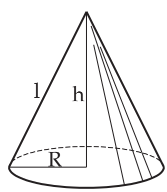


Figure 3.72

9. Right Circular Cone has a circle as base and its vertex is right above the centre of the circle **(Fig. 3.72)**. With l as the length of the apothem and R as the radius of the base

$$V = \frac{1}{3}\pi\, R^2 h, \quad (3.149) \qquad M = \pi\, R\, l = \pi\, R\sqrt{R^2 + h^2}\,, \quad (3.150) \qquad S = \pi\, R(R + l)\,. \quad (3.151)$$

10. Frustum of Right Cone or Truncated Cone (Fig. 3.73)

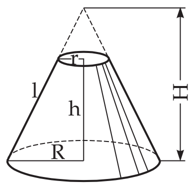


Figure 3.73

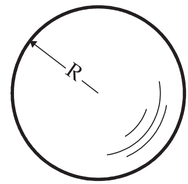


Figure 3.74

$$l = \sqrt{h^2 + (R-r)^2}, \qquad (3.152)$$

$$M = \pi\, l(R+r), \qquad (3.153)$$

$$V = \frac{\pi\, h}{3}\left(R^2 + r^2 + R\, r\right), \qquad (3.154)$$

$$H = h + \frac{h\, r}{R-r}. \qquad (3.155)$$

11. Conic Sections see 3.5.2.11, p. 206.

12. Sphere (Fig. 3.74) with radius R and diameter $D = 2R$. Every plane section of it is a circle. A plane section through the center results in a *great circle* (see 3.4.1.1, p. 160) with radius R. Only one great circle can be fitted through two surface points of the sphere if they are not the endpoints of the same diameter. The shortest connecting surface curve between two surface points is the arc of the great circle between them (see 3.4.1.1, p. 160).
Formulas for the surface and for the volume of the sphere:

$$S = 4\,\pi\, R^2 \approx 12.57\, R^2, \qquad (3.156a)$$

$$S = \pi\, D^2 \approx 3.142\, D^2, \qquad (3.156b)$$

$$S = \sqrt[3]{36\,\pi\, V^2} \approx 4.836\, \sqrt[3]{V^2}, \qquad (3.156c)$$

$$V = \frac{4}{3}\pi\, R^3 \approx 4.189\, R^3, \qquad (3.157a)$$

$$V = \frac{\pi\, D^3}{6} \approx 0.5236\, D^3, \qquad (3.157b)$$

$$V = \frac{1}{6}\sqrt{\frac{S^3}{\pi}} \approx 0.09403\, \sqrt{S^3}, \qquad (3.157c)$$

$$R = \frac{1}{2}\sqrt{\frac{S}{\pi}} \approx 0.2821\, \sqrt{S}, \qquad (3.158a)$$

$$R = \sqrt[3]{\frac{3\, V}{4\,\pi}} \approx 0.6204\, \sqrt[3]{V}. \qquad (3.158b)$$

13. Spherical Sector (Fig. 3.75)

$$S = \pi\, R(2\, h + a), \qquad (3.159)$$

$$V = \frac{2\,\pi\, R^2 h}{3}. \qquad (3.160)$$

14. Spherical Cap (Fig. 3.76)

$$a^2 = h(2\, R - h), \qquad (3.161)$$

$$V = \frac{1}{6}\pi\, h\left(3\, a^2 + h^2\right) = \frac{1}{3}\pi\, h^2(3\, R - h), \qquad (3.162)$$

$$M = 2\,\pi\, R\, h = \pi\left(a^2 + h^2\right), \qquad (3.163)$$

$$S = \pi\left(2\, R\, h + a^2\right) = \pi\left(h^2 + 2\, a^2\right). \qquad (3.164)$$

15. Spherical Layer (Fig. 3.77)

$$R^2 = a^2 + \left(\frac{a^2 - b^2 - h^2}{2\, h}\right)^2, \qquad (3.165)$$

$$V = \frac{1}{6}\pi\, h\left(3\, a^2 + 3\, b^2 + h^2\right), \qquad (3.166)$$

$$M = 2\,\pi\, R\, h, \qquad (3.167)$$

$$S = \pi\left(2\, R\, h + a^2 + b^2\right). \qquad (3.168)$$

If V_1 is the volume of a truncated cone written in a spherical layer **(Fig. 3.78)** and l is the length of its

apothem, then

$$V - V_1 = \frac{1}{6}\pi\, h\, l^2. \tag{3.169}$$

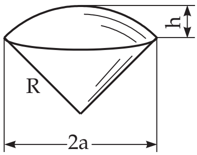


Figure 3.75

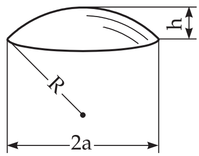


Figure 3.76

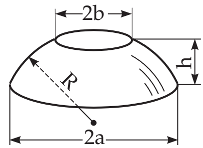


Figure 3.77

16. Torus (Fig. 3.79) is the solid which can be generated by rotating a circle around an axis which is in the plane of the circle but does not intersect it.

$$S = 4\pi^2\, R\, r \approx 39.48\, R\, r, \tag{3.170a}$$

$$S = \pi^2\, D\, d \approx 9.870\, D\, d, \tag{3.170b}$$

$$V = 2\pi^2\, R\, r^2 \approx 19.74\, R\, r^2, \tag{3.171a}$$

$$V = \frac{1}{4}\pi^2 D\, d^2 \approx 2.467\, D\, d^2. \tag{3.171b}$$

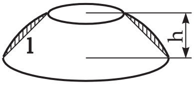


Figure 3.78

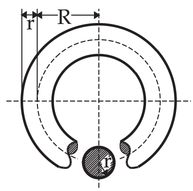


Figure 3.79

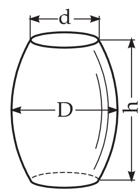


Figure 3.80

17. Barrel (Fig. 3.80) arises by rotation of a generating curve; a *circular barrel* by rotation of a circular segment, a *parabolic barrel* by rotation of a parabolic segment. For the circular barrel the following approximation formulas hold,

$$V \approx 0.262\, h \left(2\, D^2 + d^2\right) \tag{3.172a}$$

or

$$V \approx 0.0873\, h (2\, D + d)^2, \tag{3.172b}$$

and for the parabolic barrel

$$V = \frac{\pi\, h}{15}\left(2\, D^2 + D\, d + \frac{3}{4} d^2\right) \approx 0.05236 h \left(8\, D^2 + 4\, D\, d + 3\, d^2\right). \tag{3.173}$$

3.4 Spherical Trigonometry

For geodesic measures which extend over great distances the spherical shape of the Earth is taken into consideration. Therefore the spherical geometry is necessary. In particular one needs formulas for spherical triangles, i.e., triangles lying on a sphere. This was also realized by the ancient Greeks, so besides the trigonometry of the plane the trigonometry of the sphere has been developed, and nowadays Hipparchus (around 150 BC) is considered as the founder of spherical geometry.

3.4.1 Basic Concepts of Geometry on the Sphere

3.4.1.1 Curve, Arc, and Angle on the Sphere

1. Spherical Curves, Great Circle and Small Circle

Curves on the surface of a sphere are called *spherical curves*. Important spherical curves are great circles and small circles. They are *intersection circles* of a plane passing through the sphere, the so-called *intersecting plane* **(Fig. 3.81)**:
If a sphere of radius R is intersected by a plane K at distance h from the center O of the sphere, then for radius r of the intersection circle holds

$$r = \sqrt{R^2 - h^2} \quad (0 \le h \le R). \tag{3.174}$$

For $h = 0$ the intersecting plane goes through the center of the sphere, and r takes the greatest possible value. In this case the intersection circle g in the plane Γ is called a *great circle*. Every other intersection circle, with $0 < h < R$, is called a *small circle*, for instance the circle k in **Fig. 3.81**. For $h = R$ the plane K has only one common point with the sphere. Then it is called a *tangent plane*.

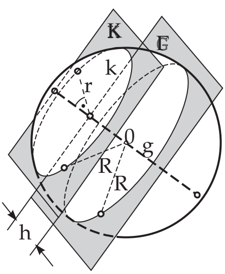


Figure 3.81

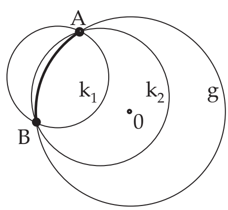


Figure 3.82

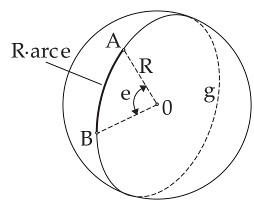


Figure 3.83

■ On the Earth the *equator* and the *meridians* with their countermeridians – which are their reflections with respect to the Earth's axis – represent great circles. The parallels of latitude are small circles (see also 3.4.1.2, p. 162).

2. Spherical Distance

Through two points A and B of the surface of the sphere, which are not opposite points, i.e., they are not the endpoints of the same diameter, infinitely many small circles can be drawn, but only one great circle (with the plane of the great circle g). Consider two small circles k_1, k_2 through A and B and turn them into the plane of the great circle passing through A and B **(Fig. 3.82)**. The great circle has the greatest radius and so the smallest curvature. So the shorter arc of the great circle is the shortest connection between A and B. It is the shortest connection between A and B on the surface of the sphere, and it is called the *spherical distance*.

3. Geodesic Lines

Geodesic lines are the curves on a surface which are the shortest connections between two points of the surface (see 3.6.3.6, p. 268).

■ In the plane the straight lines, on the sphere the great circles, are the geodesic lines (see also 3.4.1.2, p. 162).

4. Measurement of the Spherical Distance

The spherical distance of two points can be expressed as a measure of length or as a measure of angle **(Fig. 3.83)**.

1. Spherical Distance as a Measure of Angle is the angle between the radii $\overline{0A}$ and $\overline{0B}$ measured at the center 0. This angle determines the spherical distance uniquely, and hereafter it is denoted by a lowercase Latin letter. The notation can be given at the center or on the great circle arc.

2. Spherical Distance as a Measure of Length is the length of the great circle arc between A and B. It is denoted by $\overset{\frown}{AB}$ (arc AB).

3. Conversions from Measure of Angle into Measure of Length and conversely can be done by the formulas

$$\overset{\frown}{AB} = R \operatorname{arc} e = R\frac{e}{\varrho}, \quad (3.175a) \qquad e = \overset{\frown}{AB}\,\frac{\varrho}{R}. \quad (3.175b)$$

Here e denotes the angle given in degrees and arc e denotes the angle in radian (see radian measure 3.1.1.5, p. 131). The conversion factor ϱ is equal to

$$\varrho = 1\,\text{rad} = \frac{180^\circ}{\pi} = 57.2958^\circ = 3438' = 206265''. \quad (3.175c)$$

The determinations of the distance as a measure of length or angle are equivalent but in spherical trigonometry the spherical distances are given mostly as a measure of angle.

■ **A:** For spherical calculations on the Earth's surface usually a sphere is considered with the same volume as the biaxial reference ellipsoid of Krassowski. This radius of the Earth is $R = 6371.110\,\text{km}$, and consequently holds $1^\circ \mathrel{\hat{=}} 111.2\,\text{km}$, $1' \mathrel{\hat{=}} 1853.3\,\text{m} = 1$ oldseamile. Today 1 seamile = 1852 m.

■ **B:** The spherical distance between Dresden and St. Petersburg is $\overset{\frown}{AB} = 1433\,\text{km}$ or $e = \frac{1433\,\text{km}}{6371\,\text{km}}\,57.3^\circ = 12.89^\circ = 12^\circ 53'$.

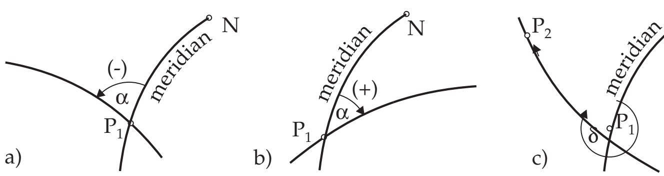


Figure 3.84

5. Intersection Angle, Course Angle, Azimuth

The *intersection angle* between spherical curves is the angle between their tangent lines at the intersection point P_1. If one of them is a meridian, the intersection angle with the curve segment to the north from P_1 is called the *course angle* α in navigation. To distinguish the inclination of the curve to the east or to the west, a sign is to be assigned to the course angle according to **Fig. 3.84a,b** restricting it to the interval $-90^\circ < \alpha \le 90^\circ$. The course angle is an oriented angle, i.e., it has a sign. It is independent of the orientation of the curve – of its sense.

The orientation of the curve from P_1 to P_2, as in **Fig. 3.84c**, can be described by the *azimuth* δ: It is the intersection angle between the northern part of the meridian passing through P_1 and the curve arc from P_1 to P_2. The azimuth is restricted to the interval $0^\circ \le \delta < 360^\circ$.

Remark: In navigation the position coordinates are usually given in sexagesimal degrees; the spherical distances and also the course angles and the azimuth are given in decimal degrees.

3.4.1.2 Special Coordinate Systems

1. Geographical Coordinates

To determine points P on the Earth's surface *geographical coordinates* are in use **(Fig. 3.85)**, i.e., spherical coordinates with the radius of the Earth, the *geographical longitude* λ and the *geographical latitude* φ.
To determine the degree of longitude the surface of the Earth is subdivided by half great circles from the north pole to the south pole, by so-called *meridians*. The zero meridian goes through the observatory of *Greenwich*. From here one counts the east longitude with the help of 180 meridians and the west longitude with 180 meridians. At the equator they are at a distance of 111 km of each other. East longitudes are given as positive, west longitudes are given as negative values. So $-180° < \lambda \le 180°$.

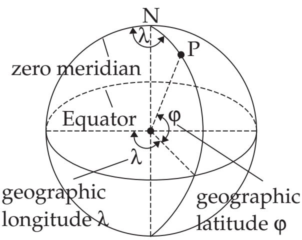


Figure 3.85

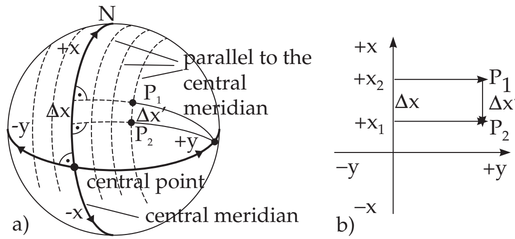


Figure 3.86

To determine the degree of the latitude the Earth's surface is divided by small circles parallel to the equator. Starting from the equator 90 degrees of latitude are to be counted to the north, the northings, and 90 southern latitudes. Northings are positive, southern latitudes are negative. So $-90° \le \varphi \le 90°$.

2. Soldner Coordinates

The right-angled *Soldner coordinates* and *Gauss-Krüger coordinates* are important in wide surface surveys. To map parts of the curved Earth's surface onto a right-angled coordinate system in a plane, distance preserving in the ordinate direction, according to Soldner the x-axis is to be placed on a meridian (it is called a central meridian), and the origin at a well-measured center point **(Fig. 3.86a)**. The ordinate y of a point P is the segment between P and the foot of the spherical orthogonal (great circle) on the central meridian. The abscissa x of the point P is the segment of a circle between P and the main parallel passing through the center, where the circle is in a parallel plane to the central meridian **(Fig. 3.86b)**.
Transferring the spherical abscissae and ordinates into the plane coordinate system the segment Δx is stretched and the directions are distorted. The *coefficient of elongation* a in the direction of the abscissa is

$$a = \frac{\Delta x}{\Delta x'} = 1 + \frac{y^2}{2R^2}, \quad R = 6371\,\text{km}. \tag{3.176}$$

To moderate the stretching, the system may not be extended more than 64 km on both sides of the central meridian. A segment of 1 km length has an elongation of 0.05 m at $y = 64$ km.

3. Gauss-Krüger Coordinates

In order to map parts of the curved Earth's surface onto the plane with an angle-preserving (conformal) mapping, at *Gauss-Krüger system* first a partition into meridian zones is prepared. For Germany these mid-meridians are at $6°, 9°, 12°$, and $15°$ east longitude **(Fig. 3.87a)**. The origin of every meridian zone is at the intersection point of the mid-meridian and the equator. In the north–south direction the total range is to be considered, in the east–west direction a $1°40'$ wide strip on both sides. In Germany it amounts ca. ± 100 km. The overlap is $20'$, which is here nearly 20 km.

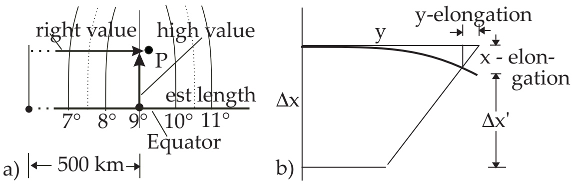


Figure 3.87

The coefficient of elongation a in the abscissa direction **(Fig. 3.87b)** is the same as in the Soldner system (3.176). To keep the mapping angle-preserving, the quantity b must be added to the ordinates:

$$b = \frac{y^3}{6\,R^2}\,. \qquad (3.177)$$

3.4.1.3 Spherical Lune or Biangle

Suppose there are two planes Γ_1 and Γ_2 passing through the endpoints A and B of a diameter of the sphere and enclosing an angle α **(Fig. 3.88)** and so defining two great circles g_1 and g_2. The part of the surface of the sphere bounded by the halves of the great circles is called a *spherical lune* or *biangle* or *spherical digon*. The sides of the spherical biangle are defined by the spherical distances between A and B on the great circles. Both are 180°.

As the angles of the biangle one defines the angles between the tangents of the great circles g_1 and g_2 at the points A and B. They are the same as the so-called *dihedral angle* α between the planes Γ_1 and Γ_2. If C and D are the bisecting points of both great circle arcs A and B , the angle α can be expressed as the spherical distance of C and D. The area A_b of the spherical biangle is proportional to the surface area of the sphere just as the angle α to 360° . Therefore the area is

$$A_\mathrm{b} = \frac{4\pi\,R^2\alpha}{360°} = \frac{2R^2\alpha}{\varrho} = 2\,R^2\,\mathrm{arc}\,\alpha \quad \text{with the conversion factor } \varrho \text{ as in (3.175c).} \qquad (3.178)$$

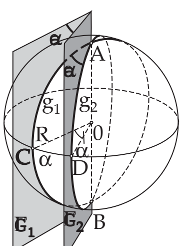


Figure 3.88

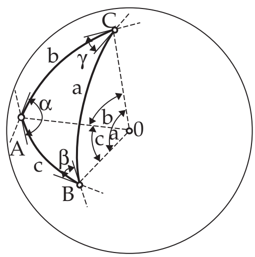


Figure 3.89

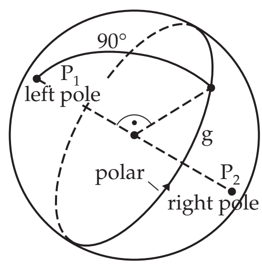


Figure 3.90

3.4.1.4 Spherical Triangle

Consider three points A, B, and C on the surface of a sphere, not on the same great circle. Connecting every two of them by a great circle yields a *spherical triangle ABC* **(Fig. 3.89)**.

The sides of the triangle are defined as the spherical distances of the points, i.e., they represent the angles at the center between the radii $\overline{0A}$, $\overline{0B}$, and $\overline{0C}$. They are denoted by a, b, and c, and hereafter they are given in angle measure, independently of whether they are denoted at the center as angles or on the surface as great circle arcs. The angles of the spherical triangle are the angles between every two planes of the great circles. They are denoted by α, β, and γ.

The order of the notation of the points, sides, and angles of the spherical triangle follows the same scheme as for triangles of the plane. A spherical triangle is called a right-sided triangle if at least one

side is equal to 90°. There is an analogy with the right-angled triangles of the plane.

3.4.1.5 Polar Triangle

1. Poles and Polar The endpoints of a diameter P_1 and P_2 are called *poles*, and the great circle g being perpendicular to this diameter is called *polar* **(Fig. 3.90)**. The spherical distance between a pole and any point of the great circle g is 90°. The orientation of the polar is defined arbitrarily: Traversing the polar along the chosen direction there is a *left pole* to the left and a *right pole* to the right.

2. Polar Triangle $A'B'C'$ of a given spherical triangle ABC is a spherical triangle such that the vertices of the original triangle are poles for its sides **(Fig. 3.91)**. For every spherical triangle ABC there exists one polar triangle $A'B'C'$. If the triangle $A'B'C'$ is the polar triangle of the spherical triangle ABC, then the triangle ABC is the polar triangle of the triangle $A'B'C'$. The angles of a spherical triangle and the corresponding sides of its polar triangle are supplementary angles, and the sides of the spherical triangle and the corresponding angles of its polar triangle are supplementary angles:

$$a' = 180° - \alpha, \qquad b' = 180° - \beta, \qquad c' = 180° - \gamma, \tag{3.179a}$$

$$\alpha' = 180° - a, \qquad \beta' = 180° - b, \qquad \gamma' = 180° - c. \tag{3.179b}$$

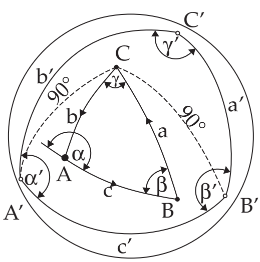
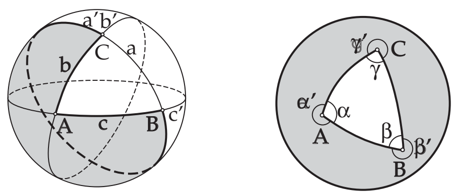


Figure 3.91



a) b)

Figure 3.92

3.4.1.6 Euler Triangles and Non-Euler Triangles

The vertices A, B, C of a spherical triangle divide every great circle into two, usually different parts. Consequently there are several different triangles with the same vertices, e.g., also the triangle with sides a', b, c and the shadowed surface in **Fig. 3.92a**. According to the definition of Euler one should always choose the arc which is smaller than 180° as a side of the spherical triangle. This corresponds to the definition of the sides as spherical distances between the vertices. Considering this, all spherical triangles of whose sides and angles are less than 180° are called *Euler triangles*, otherwise they are called non-Euler triangles. In **Fig. 3.92b** there is an Euler triangle and a *non-Euler triangle*.

3.4.1.7 Trihedral Angle

This is a three-sided solid formed by three edges s_a, s_b, s_c starting at a vertex O **(Fig. 3.93a)**. The angles a, b, c are defined as sides of the trihedral angle, every of them is enclosed by two edges. The regions between two edges are called the faces of the trihedral angle. The angles of the trihedral angle are α, β, and γ, the angles between the faces. If the vertex of a trihedral angle is at the center O of a sphere, it cuts out a spherical triangle of the surface **(Fig. 3.93b)**. The sides and the angles of the spherical triangle and the corresponding trihedral angle are coincident, so every theorem derived for a trihedral angle is valid for the corresponding spherical triangle, and conversely.

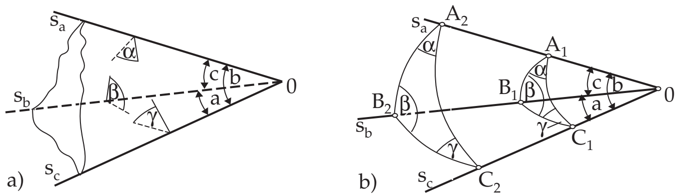


Figure 3.93

3.4.2 Basic Properties of Spherical Triangles

3.4.2.1 General Statements

For an Euler triangle with sides a, b, c, whose opposite angles are α, β, γ, the following statements are valid:

1. Sum of the Sides The sum of the sides is between 0° and 360°:

$$0^\circ < a + b + c < 360^\circ. \tag{3.180}$$

2. Sum of two Sides The sum of two sides is greater than the third one, e.g.,

$$a + b > c. \tag{3.181}$$

3. Difference of Two Sides The absolute value of the difference of two sides is smaller than the third one, e.g.,

$$|a - b| < c. \tag{3.182}$$

4. Sum of the Angles The sum of the angles is between 180° and 540°:

$$180^\circ < \alpha + \beta + \gamma < 540^\circ. \tag{3.183}$$

5. Spherical Excess The difference

$$\epsilon = \alpha + \beta + \gamma - 180^\circ \tag{3.184}$$

is called the *spherical excess*.

6. Sum of Two Angles The sum of two angles is less than the third one increased by 180°, e.g.,

$$\alpha + \beta < \gamma + 180^\circ. \tag{3.185}$$

7. Opposite Sides and Angles Opposite to a greater side there is a greater angle, and conversely.

8. Area The area A_T of a spherical triangle can be expressed by the spherical excess ϵ and by the radius of the sphere R with the formula

$$A_\mathrm{T} = \epsilon R^2 \cdot \frac{\pi}{180^\circ} = \frac{R^2 \epsilon}{\varrho} = R^2 \operatorname{arc} \epsilon. \tag{3.186a}$$

Here ϱ is the conversion factor (3.175c). From the theorem of Girard, with A_S as the surface area of the sphere, holds

$$A_\mathrm{T} = \frac{A_\mathrm{S}}{720^\circ} \epsilon. \tag{3.186b}$$

If the sides are known and not the excess, then ϵ can be calculated by the formula of L'Huilier (3.201).

3.4.2.2 Fundamental Formulas and Applications

The notation for the quantities of this paragraph corresponds to those of **Fig. 3.89**.

1. Sine Law

$$\frac{\sin a}{\sin b} = \frac{\sin \alpha}{\sin \beta}, \quad (3.187a) \qquad \frac{\sin b}{\sin c} = \frac{\sin \beta}{\sin \gamma}, \quad (3.187b) \qquad \frac{\sin c}{\sin a} = \frac{\sin \gamma}{\sin \alpha}. \quad (3.187c)$$

The equations from (3.187a) to (3.187c) can also be written as proportions, i.e., in a spherical triangle the sines of the sides are related as the sines of the opposite angles:

$$\frac{\sin a}{\sin \alpha} = \frac{\sin b}{\sin \beta} = \frac{\sin c}{\sin \gamma}. \quad (3.187d)$$

The sine law of spherical trigonometry corresponds to the sine law of plane trigonometry.

2. Cosine Law, or Cosine Law for Sides

$$\cos a = \cos b \cos c + \sin b \sin c \cos \alpha, \quad (3.188a) \qquad \cos b = \cos c \cos a + \sin c \sin a \cos \beta, \quad (3.188b)$$

$$\cos c = \cos a \cos b + \sin a \sin b \cos \gamma. \quad (3.188c)$$

The cosine law for sides in spherical trigonometry corresponds to the cosine law of plane trigonometry. From the notation one can see that the cosine law contains the three sides of the spherical triangle.

3. Sine-Cosine Law

$$\sin a \cos \beta = \cos b \sin c - \sin b \cos c \cos \alpha, \quad (3.189a)$$

$$\sin a \cos \gamma = \cos c \sin b - \sin c \cos b \cos \alpha. \quad (3.189b)$$

One can get four more equations by cyclic change of the quantities **(Fig. 3.34)**.
The sine-cosine law corresponds to the projection rule of plane trigonometry. Because it contains five quantities of the spherical triangle it is not used directly for solving problems of spherical triangles, but it is used for the derivation of further equations.

4. Cosine Law for Angles of a Spherical Triangle

$$\cos \alpha = -\cos \beta \cos \gamma + \sin \beta \sin \gamma \cos a, \quad (3.190a)$$

$$\cos \beta = -\cos \gamma \cos \alpha + \sin \gamma \sin \alpha \cos b, \quad (3.190b)$$

$$\cos \gamma = -\cos \alpha \cos \beta + \sin \alpha \sin \beta \cos c. \quad (3.190c)$$

This cosine rule contains the three angles of the spherical triangle and one of the sides. With this law one can easily express an angle by the opposite side with the angles on it, or a side by the angles; consequently every side can be expressed by the angles. Contrary to this, for plane triangles the third angle is calculated from the sum of 180°.

Remark: It is not possible to determine any side of a plane triangle from the angles, because there are infinitely many similar triangles.

5. Polar Sine-Cosine Law

$$\sin \alpha \cos b = \cos \beta \sin \gamma + \sin \beta \cos \gamma \cos a, \quad (3.191a)$$

$$\sin \alpha \cos c = \cos \gamma \sin \beta + \sin \gamma \cos \beta \cos a. \quad (3.191b)$$

Four more equations can be get by cyclic change of the quantities **(Fig. 3.34)**.
Just as for the cosine law for angles, also the polar sine-cosine law is not usually used for direct calculations for spherical triangles, but to derive further formulas.

6. Half-Angle Formulas

To determine an angle of a spherical triangle from the sides one can use the cosine law for sides. The half-angle formulas allow us to calculate the angles by their tangents, similarly to the half-angle formulas of plane trigonometry:

$$\tan \frac{\alpha}{2} = \sqrt{\frac{\sin(s-b)\sin(s-c)}{\sin s \sin(s-a)}}, \quad (3.192a) \qquad \tan \frac{\beta}{2} = \sqrt{\frac{\sin(s-c)\sin(s-a)}{\sin s \sin(s-b)}}, \quad (3.192b)$$

$$\tan\frac{\gamma}{2} = \sqrt{\frac{\sin(s-a)\sin(s-b)}{\sin s\sin(s-c)}}, \quad (3.192c) \qquad s = \frac{a+b+c}{2}. \quad (3.192d)$$

If from the three sides of a spherical triangle all the three angles should be determined, then the following calculations are useful:

$$\tan\frac{\alpha}{2} = \frac{k}{\sin(s-a)}, \quad (3.193a) \qquad \tan\frac{\beta}{2} = \frac{k}{\sin(s-b)}, \quad (3.193b)$$

$$\tan\frac{\gamma}{2} = \frac{k}{\sin(s-c)} \quad \text{with} \quad (3.193c)$$

$$k = \sqrt{\frac{\sin(s-a)\sin(s-b)\sin(s-c)}{\sin s}}, \quad (3.193d) \qquad s = \frac{a+b+c}{2}. \quad (3.193e)$$

7. Half-Side Formulas

With the half-side formulas it is possible to determine one side or all three sides of a spherical triangle from its three angles:

$$\cot\frac{a}{2} = \sqrt{\frac{\cos(\sigma-\beta)\cos(\sigma-\gamma)}{-\cos\sigma\cos(\sigma-\alpha)}}, \quad (3.194a) \qquad \cot\frac{b}{2} = \sqrt{\frac{\cos(\sigma-\gamma)\cos(\sigma-\alpha)}{-\cos\sigma\cos(\sigma-\beta)}}, \quad (3.194b)$$

$$\cot\frac{c}{2} = \sqrt{\frac{\cos(\sigma-\alpha)\cos(\sigma-\beta)}{-\cos\sigma\cos(\sigma-\gamma)}}, \quad (3.194c) \qquad \sigma = \frac{\alpha+\beta+\gamma}{2}, \quad (3.194d)$$

or

$$\cot\frac{a}{2} = \frac{k'}{\cos(\sigma-\alpha)}, \quad (3.195a) \qquad \cot\frac{b}{2} = \frac{k'}{\cos(\sigma-\beta)}, \quad (3.195b)$$

$$\cot\frac{c}{2} = \frac{k'}{\cos(\sigma-\gamma)} \quad \text{with} \quad (3.195c)$$

$$k' = \sqrt{\frac{\cos(\sigma-\alpha)\cos(\sigma-\beta)\cos(\sigma-\gamma)}{-\cos\sigma}}, \quad (3.195d) \qquad \sigma = \frac{\alpha+\beta+\gamma}{2}. \quad (3.195e)$$

Since for the sum of the angles of a spherical triangle according to (3.183):

$$180^\circ < 2\sigma < 540^\circ \quad \text{or} \quad 90^\circ < \sigma < 270^\circ \quad (3.196)$$

holds, $\cos\sigma < 0$ must always be valid. Because of the requirements for Euler triangles all the roots are real.

8. Applications of the Fundamental Formulas of Spherical Geometry

With the help of the given fundamental formulas, for instance distances, the azimuth, and course angles on the Earth can be determined.

■ **A:** Determine the shortest distance between Dresden ($\lambda_1 = 13°46', \varphi_1 = 51°16'$) and Alma Ata ($\lambda_2 = 76°55', \varphi_2 = 43°18'$).

Solution: The geographical coordinates (λ_1, φ_1), (λ_2, φ_2) and the north pole N **(Fig. 3.94)** result in two sides of the triangle P_1P_2N $a = 90° - \varphi_2$ and $b = 90° - \varphi_1$ lying on meridians, and also the angle between them $\gamma = \lambda_2 - \lambda_1$. For $c = e$ it follows from the cosine law (3.188a)

$$\cos c = \cos a\cos b + \sin a\sin b\cos\gamma,$$

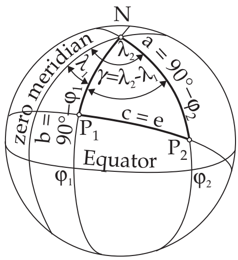


Figure 3.94

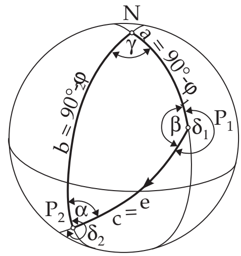


Figure 3.95

$$\begin{aligned}\cos e &= \cos(90° - \varphi_1)\cos(90° - \varphi_2) + \sin(90° - \varphi_1)\sin(90° - \varphi_2)\cos(\lambda_2 - \lambda_1)\\ &= \sin\varphi_1 \sin\varphi_2 + \cos\varphi_1 \cos\varphi_2 \cos(\lambda_2 - \lambda_1),\end{aligned} \quad (3.197)$$

i.e., $\cos e = 0.53498 + 0.20567 = 0.74065, e = 42.213°$. The great circle segment $\widehat{P_1P_2}$ has length 4694 km using (3.175a).

■ **B:** Calculate the course angles δ_1 and δ_2 at departure and at arrival, and also the distance in sea miles of a voyage from Bombay ($\lambda_1 = 72°48'$, $\varphi_1 = 19°00'$) to Dar es Saalam ($\lambda_2 = 39°28'$, $\varphi_2 = -6°49'$) along a great circle.

Solution: The calculation of the two sides $a = 90° - \varphi_1 = 71°00', b = 90° - \varphi_2 = 96°49'$ and the enclosed angle $\gamma = \lambda_1 - \lambda_2 = 33°20'$ in the spherical triangle P_1P_2N with the help of the geographical coordinates $(\lambda_1, \varphi_1), (\lambda_2, \varphi_2)$ **(Fig. 3.95)** and the cosine law (3.188c) $\cos c = \cos e = \cos a \cos b + \sin a \sin b \cos\gamma$ yields $\widehat{P_1P_2} = e = 41.777°$, and because $1' \approx 1\,\text{sm}$ follows $\widehat{P_1P_2} \approx 2507\,\text{sm}$.
With the cosine law for sides (3.188a)

$$\alpha = \arccos\frac{\cos a - \cos b\cos c}{\sin b \sin c} = 51.248° \quad \text{and} \quad \beta = \arccos\frac{\cos b - \cos a \cos c}{\sin a \sin c} = 125.018°.$$

Therefore, the results are

$$\delta_1 = 360° - \beta = 234.982° \quad \text{and} \quad \delta_2 = 180° + \alpha = 231.248°.$$

Remark: It makes sense to use the sine law to determine sides and angles only if it is already obvious from the problem that the angles are acute or obtuse.

3.4.2.3 Further Formulas

1. Delambre Equations

Analogously to the Mollweide formulas of plane trigonometry the corresponding formulas of Delambre are valid for spherical triangles:

$$\frac{\cos\frac{\alpha-\beta}{2}}{\sin\frac{\gamma}{2}} = \frac{\sin\frac{a+b}{2}}{\sin\frac{c}{2}}, \quad (3.198a) \qquad \frac{\sin\frac{\alpha-\beta}{2}}{\cos\frac{\gamma}{2}} = \frac{\sin\frac{a-b}{2}}{\sin\frac{c}{2}}, \quad (3.198b)$$

$$\frac{\cos\frac{\alpha+\beta}{2}}{\sin\frac{\gamma}{2}} = \frac{\cos\frac{a+b}{2}}{\cos\frac{c}{2}}, \quad (3.198c) \qquad \frac{\sin\frac{\alpha+\beta}{2}}{\cos\frac{\gamma}{2}} = \frac{\cos\frac{a-b}{2}}{\cos\frac{c}{2}}. \quad (3.198d)$$

Since for every equation two more equations exist by cyclic changes, altogether there are 12 Delambre equations.

2. Neper Equations and Tangent Law

$$\tan\frac{\alpha-\beta}{2}=\frac{\sin\frac{a-b}{2}}{\sin\frac{a+b}{2}}\cot\frac{\gamma}{2}, \quad (3.199a) \qquad \tan\frac{\alpha+\beta}{2}=\frac{\cos\frac{a-b}{2}}{\cos\frac{a+b}{2}}\cot\frac{\gamma}{2}, \quad (3.199b)$$

$$\tan\frac{a-b}{2}=\frac{\sin\frac{\alpha-\beta}{2}}{\sin\frac{\alpha+\beta}{2}}\tan\frac{c}{2}, \quad (3.199c) \qquad \tan\frac{a+b}{2}=\frac{\cos\frac{\alpha-\beta}{2}}{\cos\frac{\alpha+\beta}{2}}\tan\frac{c}{2}. \quad (3.199d)$$

These equations are also called *Neper analogies*. From these one can derive formulas analogous to the tangent law of plane trigonometry:

$$\frac{\tan\frac{a-b}{2}}{\tan\frac{a+b}{2}}=\frac{\tan\frac{\alpha-\beta}{2}}{\tan\frac{\alpha+\beta}{2}}, \quad (3.200a) \qquad \frac{\tan\frac{b-c}{2}}{\tan\frac{b+c}{2}}=\frac{\tan\frac{\beta-\gamma}{2}}{\tan\frac{\beta+\gamma}{2}}, \quad (3.200b)$$

$$\frac{\tan\frac{c-a}{2}}{\tan\frac{c+a}{2}}=\frac{\tan\frac{\gamma-\alpha}{2}}{\tan\frac{\gamma+\alpha}{2}}. \quad (3.200c)$$

3. L'Huilier Equations

The area of a spherical triangle can be calculated with the help of the excess ϵ, either from the known angles α, β, γ according to (3.184), or if the three sides a, b, c are known, with the formulas about the angles from (3.193a) to (3.193e). The L'Huilier equation makes possible the direct calculation of ϵ from the sides:

$$\tan\frac{\epsilon}{4}=\sqrt{\tan\frac{s}{2}\tan\frac{s-a}{2}\tan\frac{s-b}{2}\tan\frac{s-c}{2}}. \quad (3.201)$$

This equation corresponds to the Heron formula of plane trigonometry.

3.4.3 Calculation of Spherical Triangles

3.4.3.1 Basic Problems, Accuracy Observations

The different cases occurring most often in calculations of spherical triangles are arranged into so-called *basic problems*. For every basic problem for acute-angled spherical triangles there are several ways to solve it, and it depends on whether the calculations are based only on the formulas from (3.187a) to (3.191b) or also on the formulas from (3.192a) to (3.201), and also whether one is looking for only one or more quantities of the triangle.

Formulas containing the tangent function yield numerically more accurate results, especially in comparison to the determination of defining quantities with the sine function if they are close to 90°, and with the cosine function if their value is close to 0° or 180°. For Euler triangles the quantities calculated with the sine law are bi-valued, since the sine function is positive in both of the first quadrants, while the results obtained from other functions are unique.

3.4.3.2 Right-Angled Spherical Triangles

1. Special Formulas

In a right-angled spherical triangle at least one of the angles is 90°. The sides and angles are denoted analogously to the plane right-angled triangle. If as in **Fig. 3.96** γ is a right angle, the side c is called the hypotenuse, a and b the legs and α and β are the leg angles. From the equations from (3.187d) to

(3.191b) it follows for $\gamma = 90°$:

$$\sin a = \sin\alpha \sin c, \qquad (3.202a) \qquad\qquad \sin b = \sin\beta \sin c, \qquad (3.202b)$$

$$\cos c = \cos a \cos b, \qquad (3.202c) \qquad\qquad \cos c = \cot\alpha \cot\beta, \qquad (3.202d)$$

$$\tan a = \cos\beta \tan c, \qquad (3.202e) \qquad\qquad \tan b = \cos\alpha \tan c, \qquad (3.202f)$$

$$\tan b = \sin a \tan\beta, \qquad (3.202g) \qquad\qquad \tan a = \sin b \tan\alpha, \qquad (3.202h)$$

$$\cos\alpha = \sin\beta \cos a, \qquad (3.202i) \qquad\qquad \cos\beta = \sin\alpha \cos b. \qquad (3.202j)$$

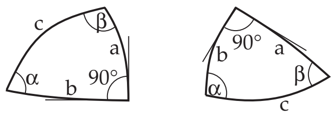


Figure 3.96

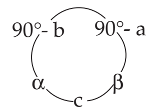


Figure 3.97

If in certain problems other sides or angles are given, for instance instead of α, β, γ the quantities b, γ, α, the necessary equations can be get by cyclic change of these quantities. For calculations in a right-angled spherical triangle one usually starts with three given quantities, the angle $\gamma = 90°$ and two other quantities. There are six basic problems, which are represented in **Table 3.8**.

Table 3.8 Defining quantities of a spherical right-angled triangle

Basic problem	**Given defining quantities**	**Number of the formula to determine other quantities**
1.	Hypotenuse and a leg c, a	α (3.202a), β (3.202e), b (3.202c)
2.	Two legs a, b	α (3.202h), β (3.202g), c (3.202c)
3.	Hypotenuse and an angle c, α	a (3.202a), b (3.202f), β (3.202d)
4.	Leg and the angles on it a, β	c (3.202e), b (3.202j), α (3.202i)
5.	Leg and the angle opposite to it a, α	b (3.202h), c (3.202a), β (3.202i)
6.	Two angles α, β	a (3.202i), b (3.202j), c (3.202d)

2. Neper Rule

The Neper rule summarizes the equations (3.202a) to (3.202j). If the five determining quantities of a right-angled spherical triangle, not considering the right angle, are arranged along a circle in the same order as they are in the triangle, and if the legs are replaced by their complementary angles $90° - a$, $90° - b$, **(Fig. 3.97)**, then the following is valid:

1. The cosine of every defining quantity is equal to the product of the cotangent values of its neighboring quantities.

2. The cosine of every defining quantity is equal to the product of the sine values of the non-neighboring quantities.

■ **A:** $\cos\alpha = \cot(90° - b)\cot c = \dfrac{\tan b}{\tan c}$ (see (3.202f) .

■ **B:** $\cos(90° - a) = \sin c \sin\alpha = \sin a$ (see (3.202a)).

■ **C:** Map the sphere given by a grid, onto a cylinder which contacts the sphere along a meridian, the

so-called *central meridian*. This meridian and the equator form the axis of the Gauss-Krüger system **(Fig. 3.98a,b)**.

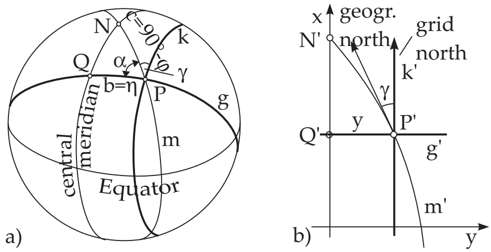


Figure 3.98

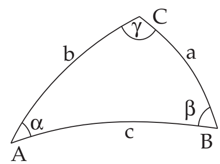


Figure 3.99

Solution: A point P of the surface of the sphere will have the corresponding point P' on the plane. The great circle g passing through the point P perpendicular to the central meridian is mapped into a line g' perpendicular to the x-axis, and the small circle k passing through P parallel to the given meridian becomes a line k' parallel to the x-axis (*grid meridian*). The image of the meridian m through P is not line but a curve m' (*true meridian*). The upward direction of the tangent of m' at P' gives the *geographical north*, the upward direction of k' gives the *grid north* direction. The angle γ between the two north directions is called the *meridian convergence*.

In a right-angled spherical triangle QPN with $c = 90° - \varphi$, and $b = \eta$ one gets γ from $\alpha = 90° - \gamma$. The Neper rule yields $\cos\alpha = \dfrac{\tan b}{\tan c}$ or $\cos(90° - \gamma) = \dfrac{\tan\eta}{\tan(90° - \varphi)}$, $\sin\gamma = \tan\eta\tan\varphi$. Because γ and η are mostly small, one can consider $\sin\gamma \approx \gamma$, $\tan\eta \approx \eta$; consequently $\gamma = \eta\tan\varphi$ is valid. The length deviation γ of this cylinder sketch is pretty small for small distances η, and so one can substitute $\eta = y/R$, where y is the ordinate of P. This yields $\gamma = (y/R)\tan\varphi$. The conversion of γ from radian into degree measure yields a meridian convergence of $\gamma = 0.018706$ or $\gamma = 1°04'19''$ at $\varphi = 50°$, $y = 100\,\text{km}$.

3.4.3.3 Spherical Triangles with Oblique Angles

For three given quantities, there is to be distinguished between six basic problems, just as it was done for right-angled spherical triangles. The notation for the angles is α, β, γ and a, b, c for the opposite sides **(Fig. 3.99)**.

Tables 3.9, 3.10, 3.11, and **3.12** summarize, which formulas should be used for which defining quantities in the case of the six basic problems. Problems 3, 4, 5, and 6 can also be solved by decomposing the general triangle into two right-angled triangles. To do this for problems 3 and 4 **(Fig. 3.100, Fig. 3.101)** one can use the *spherical perpendicular* from B to AC to the point D, and for problems 5 and 6 **(Fig. 3.102)** from C to AB to the point D.

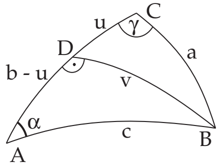


Figure 3.100

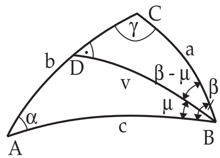


Figure 3.101

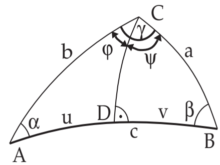


Figure 3.102

In the headings of **Tables 3.9, 3.10, 3.11**, and **3.12** the given sides and angles are denoted by S and W respectively. This means for instance SWS: Two sides and the enclosed angle are given.

Table 3.9 First and second basic problems for spherical oblique triangles

First basic problem **Given: 3 sides $\boldsymbol{a, b, c}$** SSS	**Second basic problem** **Given: 3 angles $\boldsymbol{\alpha, \beta, \gamma}$** WWW
Conditions: $0° < a+b+c < 360°$, $a+b>c,\ a+c>b,\ b+c>a$.	Conditions: $180° < \alpha+\beta+\gamma < 540°$, $\alpha+\beta < 180°+\gamma, \alpha+\gamma < 180°+\beta$, $\beta+\gamma < 180°+\alpha$.
Solution 1: Required α. $\cos\alpha = \dfrac{\cos a - \cos b\cos c}{\sin b \sin c}$ or $\tan\dfrac{\alpha}{2} = \sqrt{\dfrac{\sin(s-b)\sin(s-c)}{\sin s\sin(s-a)}}$, $s = \dfrac{a+b+c}{2}$. Solution 2: Required α, β, γ. $k = \sqrt{\dfrac{\sin(s-a)\sin(s-b)\sin(s-c)}{\sin s}}$, $\tan\dfrac{\alpha}{2} = \dfrac{k}{\sin(s-a)}$, $\tan\dfrac{\beta}{2} = \dfrac{k}{\sin(s-b)}$, $\tan\dfrac{\gamma}{2} = \dfrac{k}{\sin(s-c)}$. Checking: $(s-a)+(s-b)+(s-c) = s$, $\tan\dfrac{\alpha}{2}\tan\dfrac{\beta}{2}\tan\dfrac{\gamma}{2}\sin s = k$.	Solution 1: Required a. $\cos a = \dfrac{\cos\alpha + \cos\beta\cos\gamma}{\sin\beta\sin\gamma}$ or $\cot\dfrac{a}{2} = \sqrt{\dfrac{\cos(\sigma-\beta)\cos(\sigma-\gamma)}{-\cos\sigma\cos(\sigma-\alpha)}}$, $\sigma = \dfrac{\alpha+\beta+\gamma}{2}$. Solution 2: Required a, b, c. $k' = \sqrt{\dfrac{\cos(\sigma-\alpha)\cos(\sigma-\beta)\cos(\sigma-\gamma)}{-\cos\sigma}}$. $\cot\dfrac{a}{2} = \dfrac{k'}{\cos(\sigma-\alpha)}$, $\cot\dfrac{b}{2} = \dfrac{k'}{\cos(\sigma-\beta)}$, $\cot\dfrac{c}{2} = \dfrac{k'}{\cos(\sigma-\gamma)}$. Checking: $(\sigma-\alpha)+(\sigma-\beta)+(\sigma-\gamma) = \sigma$, $\cot\dfrac{a}{2}\cot\dfrac{b}{2}\cot\dfrac{c}{2}(-\cos\sigma) = k'$.

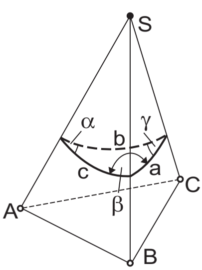


Figure 3.103

■ **A Tetrahedron:** A tetrahedron has base ABC and vertex S **(Fig. 3.103)**. The faces ABS and BCS intersect each other at an angle $\beta = 74°18'$, BCS and CAS at an angle $\gamma = 63°40'$, and CAS and ABS at $\alpha = 80°00'$. How big are the angles between every two of the edges AS, BS, and CS?
Solution: From a spherical surface around the vertex S of the pyramid, the trihedral angle **(Fig. 3.103)** cuts out a spherical triangle with sides a, b, c.
The angles between the faces are the angles of the spherical triangle, the angles between the edges, which are searched for, are the sides. The determination of the angles a, b, c corresponds to the second problem. Solution 2 in **Table 3.9** yields:
$\sigma = 108°59'$, $\sigma-\alpha = 28°59'$, $\sigma-\beta = 34°41'$, $\sigma-\gamma = 45°19'$, $k' = 1.246983$, $\cot(a/2) = 1.425514$, $\cot(b/2) = 1.516440$, $\cot(c/2) = 1.773328$.

■ **B Radio Bearing:** In the case of a radio bearing two fixed stations $P_1(\lambda_1, \varphi_1)$ and $P_2(\lambda_2, \varphi_2)$ receive the azimuths δ_1 and δ_2 via the radio waves emitted by a ship **(Fig. 3.104)**. The task is to determine the geographical coordinates of the position P_0 of the ship. The problem, known by *marines* as the *shore-to-ship bearing*, is an *intersection problem on the sphere*, and it will be solved similar to the intersection problem in the plane (see 3.2.2.3, p. 148).

1. Calculation in triangle P_1P_2N: In triangle P_1P_2N the sides $P_1N = 90° - \varphi_1$, $P_2N = 90° - \varphi_2$ and

the angle $\sphericalangle P_1NP_2 = \lambda_2 - \lambda_1 = \Delta\lambda$ are given. The calculations of the angle $\sphericalangle \varepsilon_1,\ \varepsilon_2$ and the segment $P_1P_2 = e$ correspond to the third basic problem.

Table 3.10 Third basic problem for spherical oblique triangles

Third basic problem **Given: 2 sides and the enclosed angle, e.g., a, b, γ** **SWS**	
Condition: none	
Solution 1: Required c, or c and α. $\cos c = \cos a \cos b + \sin a \sin b \cos\gamma$, $\sin\alpha = \dfrac{\sin a \sin\gamma}{\sin c}$. α can be in quadrant I. or II. We apply the theorem: Larger angle is opposite to the larger side or checking calculation: $\cos a - \cos b \cos c \gtrless 0 \rightarrow \begin{matrix}\alpha \text{ in q. I.}\\ \alpha \text{ in q. II.}\end{matrix}$ Solution 2: Required α, or α and c. $\tan u = \tan a \cos\gamma$ $\tan\alpha = \dfrac{\tan\gamma \sin u}{\sin(b-u)}$ $\tan c = \dfrac{\tan(b-u)}{\cos\alpha}$. Solution 3: Required α and (or) β. $\tan\dfrac{\alpha+\beta}{2} = \dfrac{\cos\dfrac{a-b}{2}}{\cos\dfrac{a+b}{2}}\cot\dfrac{\gamma}{2}$.	$\tan\dfrac{\alpha-\beta}{2} = \dfrac{\sin\dfrac{a-b}{2}}{\sin\dfrac{a+b}{2}}\cot\dfrac{\gamma}{2}$ $\left(-90^\circ < \dfrac{\alpha-\beta}{2} < 90^\circ\right)$ $\alpha = \dfrac{\alpha+\beta}{2} + \dfrac{\alpha-\beta}{2},\quad \beta = \dfrac{\alpha+\beta}{2} - \dfrac{\alpha-\beta}{2}$. Solution 4: Required α, β, c. $\tan\dfrac{\alpha+\beta}{2} = \dfrac{\cos\dfrac{a-b}{2}\cos\dfrac{\gamma}{2}}{\cos\dfrac{a+b}{2}\sin\dfrac{\gamma}{2}} = \dfrac{Z}{N}$, $\tan\dfrac{\alpha-\beta}{2} = \dfrac{\sin\dfrac{a-b}{2}\cos\dfrac{\gamma}{2}}{\sin\dfrac{a+b}{2}\sin\dfrac{\gamma}{2}} = \dfrac{Z'}{N'}$ $\left(-90^\circ < \dfrac{\alpha+\beta}{2} < 90^\circ\right)$ $\alpha = \dfrac{\alpha+\beta}{2} + \dfrac{\alpha-\beta}{2},\quad \beta = \dfrac{\alpha+\beta}{2} - \dfrac{\alpha-\beta}{2}$, $\cos\dfrac{c}{2} = \dfrac{Z}{\sin\dfrac{\alpha+\beta}{2}},\quad \sin\dfrac{c}{2} = \dfrac{Z'}{\sin\dfrac{\alpha-\beta}{2}}$. Checking: Double calculation of c.

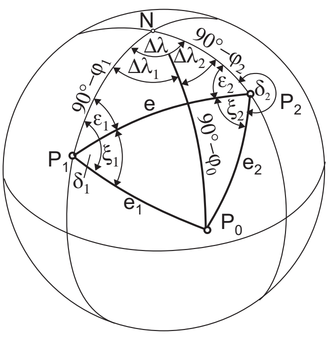


Figure 3.104

2. Calculation in the triangle $P_1P_2P_0$: Because $\xi_1 = \delta_1 - \varepsilon_1,\ \xi_2 = 360^\circ - (\delta_2 + \varepsilon_2)$, the side e and the angles lying on it ξ_1 and ξ_2 are known in $P_1P_0P_2$. Calculations of the sides e_1 and e_2 correspond to the fourth basic problem, third solution. The coordinates of the point P_0 can be calculated from the azimuth and the distance from P_1 or P_2.

3. Calculation in the triangle NP_1P_0: In the triangle NP_1P_0 there are given two sides $NP_1 = 90^\circ - \varphi_1,\ P_1P_0 = e_1$ and the included angle δ_1. The side $NP_0 = 90^\circ - \varphi_0$ and the angle $\Delta\lambda_1$ are calculated according to the third basic problem, first solution. For checking one can calculate in the triangle NP_0P_2 for a second time $NP_0 = 90^\circ - \varphi_0$, and also $\Delta\lambda_2$. Consequently, the longitude $\lambda_0 = \lambda_1 + \Delta\lambda_1 = \lambda_2 - \Delta\lambda_2$ and the latitude φ_0 of the point P_0 are known.

3.4.3.4 Spherical Curves

Spherical trigonometry has a very important application in navigation. One basic problem is to determine the course angle, which gives the optimal route. Other fields of application are geodesic surveys and also robot-movement planning.

Table 3.11 Fourth basic problem for spherical oblique triangles

<table>
<tr><td colspan="2">Fourth basic problem
Given: One side and two adjacent angles, e.g., α, β, c WSW</td></tr>
<tr><td colspan="2">Condition: none</td></tr>
<tr><td>
<u>Solution 1:</u> Required γ, or γ and a.
$\cos\gamma = -\cos\alpha\cos\beta + \sin\alpha\sin\beta\cos c\,,$
$\sin a = \frac{\sin c\sin\alpha}{\sin\gamma}\,.$
a can be in quadrant I or II. We apply the theorem: The larger side is in opposite to the larger angle or checking calculation:
$\cos\alpha + \cos\beta\cos\gamma \gtrless 0 \rightarrow \begin{matrix}\alpha \text{ in q. I.}\\ \alpha \text{ in q. II.}\end{matrix}$
<u>Solution 2:</u> Required a, or a and γ.
$\cot\mu = \tan\alpha\cos c\,,\ \tan a = \frac{\tan c\cos\mu}{\cos(\beta-\mu)}\,,$
$\tan\gamma = \frac{\cot(\beta-\mu)}{\cos a}\,.$
<u>Solution 3:</u> Required a and (or) b.
$\tan\frac{a+b}{2} = \frac{\cos\frac{\alpha-\beta}{2}}{\cos\frac{\alpha+\beta}{2}}\tan\frac{c}{2}\,,$
$\tan\frac{a-b}{2} = \frac{\sin\frac{\alpha-\beta}{2}}{\sin\frac{\alpha+\beta}{2}}\tan\frac{c}{2}$
</td><td>
$\left(-90^\circ < \frac{a-b}{2} < 90^\circ\right),$
$a = \frac{a+b}{2} + \frac{a-b}{2}\,,\quad b = \frac{a+b}{2} - \frac{a-b}{2}\,.$
<u>Solution 4:</u> Required a, b, γ.
$\tan\frac{a+b}{2} = \frac{\cos\frac{\alpha-\beta}{2}\sin\frac{c}{2}}{\cos\frac{\alpha+\beta}{2}\cos\frac{c}{2}} = \frac{Z}{N}\,,$
$\tan\frac{a-b}{2} = \frac{\sin\frac{\alpha-\beta}{2}\sin\frac{c}{2}}{\sin\frac{\alpha+\beta}{2}\cos\frac{c}{2}} = \frac{Z'}{N'}$
$\left(90^\circ < \frac{a-b}{2} < 90^\circ\right),$
$a = \frac{a+b}{2} + \frac{a-b}{2}\,,\quad b = \frac{a+b}{2} - \frac{a-b}{2}\,,$
$\sin\frac{\gamma}{2} = \frac{Z}{\sin\frac{a+b}{2}}\,,\quad \cos\frac{\gamma}{2} = \frac{Z'}{\sin\frac{a-b}{2}}\,.$
Checking: Double calculation of γ.
</td></tr>
</table>

1. Orthodrome

1. Notion The geodesic lines of the surface of the sphere – which are curves, connecting two points A and B by the shortest path – are called *orthodromes* or *great circles* (see 3.4.1.1, **3.**, p. 160).

2. Equation of the Orthodrome Moving on an orthodrome – except for meridians and the equator – needs a continuous change of course angle. These orthodromes with position-dependent course angles α can be given uniquely by their point closest to the north pole $P_{\mathrm{N}}(\lambda_{\mathrm{N}},\,\varphi_{\mathrm{N}})$, where $\varphi_{\mathrm{N}} > 0^\circ$ holds. The orthodrome has the course angle $\alpha_{\mathrm{N}} = 90^\circ$ at the point closest to the north pole. The equation of the orthodrome through P_{N} and the running point $Q(\lambda, \varphi)$, whose relative position to P_{N} is arbitrary, can be given by the Neper rule (3.4.3.2,**2.**, p. 170) according to **Fig. 3.105** as:

$$\tan\varphi_{\mathrm{N}}\cos(\lambda-\lambda_{\mathrm{N}}) = \tan\varphi. \tag{3.203}$$

Point Closest to the North Pole: The coordinates of the point closest to the north pole $P_{\mathrm{N}}(\lambda_{\mathrm{N}},\,\varphi_{\mathrm{N}})$ of an orthodrome with a course angle α_A $(\alpha_A \neq 0^\circ)$ at the point $A(\lambda_A,\,\varphi_A)$ $(\varphi_A \neq 90^\circ)$ can be calculated by the Neper rule (3.4.3.2,**2.**, p. 170) considering the relative position of P_{N} and the sign of α_A according to **Fig. 3.106** as:

$$\varphi_{\mathrm{N}} = \arccos(\sin|\alpha_A|\cos\varphi_A) \quad (3.204\mathrm{a}) \quad \text{and} \quad \lambda_{\mathrm{N}} = \lambda_A + \operatorname{sign}(\alpha_A)\left|\arccos\frac{\tan\varphi_A}{\tan\varphi_{\mathrm{N}}}\right|. \quad (3.204\mathrm{b})$$

Remark: If a calculated geographical distance λ is not in the domain $-180° < \lambda \le 180°$, then for $\lambda \ne \pm k \cdot 180°$ $(k \in \mathbb{N})$ the *reduced geographical distance* λ_{red} is:

$$\lambda_{\text{red}} = 2 \arctan\left(\tan\frac{\lambda}{2}\right). \qquad (3.205)$$

This is called the *reduction* of the angle in the domain.

Table 3.12 Fifth and sixth basic problems for a spherical oblique triangle

Fifth basic problem SSW **Given: 2 sides and an angle opposite to one of them, e.g., a, b, α**	**Sixth basic problem** WWS **Given: 2 angles and a side opposite to one of them, e.g., a, α, β**
Conditions: See distinction of cases.	Conditions: See distinction of cases.
Solution: Required is any missing quantity. $\sin\beta = \frac{\sin b \sin\alpha}{\sin a}$ 2 values β_1, β_2 are possible. Let β_1 be acute and $\beta_2 = 180° - \beta_1$ obtuse. Distinction of cases: 1. $\frac{\sin b \sin\alpha}{\sin a} > 1$ 0 solution. 2. $\frac{\sin b \sin\alpha}{\sin a} = 1$ 1 solution $\beta = 90°$. 3. $\frac{\sin b \sin\alpha}{\sin a} < 1$: 3.1. $\sin a > \sin b$: 3.1.1. $b < 90°$ 1 solution β_1. 3.1.2. $b > 90°$ 1 solution β_2. 3.2. $\sin a < \sin b$: 3.2.1. $a < 90°, \alpha < 90°$; $a > 90°, \alpha > 90°$ } 2 solutions β_1, β_2. 3.2.2. $a < 90°, \alpha > 90°$; $a > 90°, \alpha < 90°$ } 0 solution. Further calculations with one angle or with two angles β:	Solution: Required is any missing quantity. $\sin b = \frac{\sin a \sin\beta}{\sin\alpha}$ 2 values b_1, b_2 are possible. Let b_1 be acute and $b_2 = 180° - \beta_1$ obtuse. Distinction of cases: 1. $\frac{\sin a \sin\beta}{\sin\alpha} > 1$ 0 solution. 2. $\frac{\sin a \sin\beta}{\sin\alpha} = 1$ 1 solution $b = 90°$. 3. $\frac{\sin a \sin\beta}{\sin\alpha} < 1$: 3.1. $\sin\alpha > \sin\beta$. 3.1.1. $\beta < 90°$ 1 solution b_1. 3.1.2. $\beta > 90°$ 1 solution b_2. 3.2. $\sin\alpha < \sin\beta$: 3.2.1. $a < 90°, \alpha < 90°$; $a > 90°, \alpha > 90°$ } 2 solutions b_1, b_2. 3.2.2. $a < 90°, \alpha > 90°$; $a > 90°, \alpha < 90°$ } 0 solution. Further calculations with one side or with two sides b:

Method 1:	Method 2:	
$\tan u = \tan b \cos\alpha$, $\tan v = \tan a \cos\beta$, $c = u + v$, $\cot\varphi = \cos b \tan\alpha$, $\cot\psi = \cos a \tan\beta$, $\gamma = \varphi + \psi$.	$\tan\frac{c}{2} = \tan\frac{a+b}{2}\frac{\cos\frac{\alpha+\beta}{2}}{\cos\frac{\alpha-\beta}{2}}$, $= \tan\frac{a-b}{2}\frac{\sin\frac{\alpha+\beta}{2}}{\sin\frac{\alpha-\beta}{2}}$.	$\tan\frac{\gamma}{2} = \cot\frac{\alpha+\beta}{2}\frac{\cos\frac{a-b}{2}}{\sin\frac{a+b}{2}}$, $= \cot\frac{\alpha-\beta}{2}\frac{\sin\frac{a-b}{2}}{\sin\frac{a+b}{2}}$.
Checking: Double calculation of $\frac{c}{2}$ and $\frac{\gamma}{2}$		

Intersection Points with the Equator: The intersection points $P_{E_1}(\lambda_{E_1}, 0°)$ and $P_{E_2}(\lambda_{E_2}, 0°)$ of the orthodrome with the equator can be calculated from (3.203) because $\tan\varphi_N \cos(\lambda_{E_\nu} - \lambda_N) = 0$ $(\nu = 1, 2)$ must hold:

$$\lambda_{E_\nu} = \lambda_N \mp 90° \qquad (\nu = 1, 2). \qquad (3.206)$$

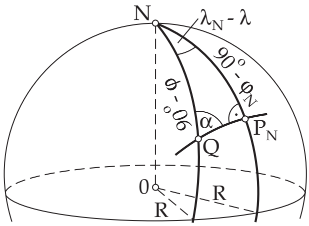


Figure 3.105

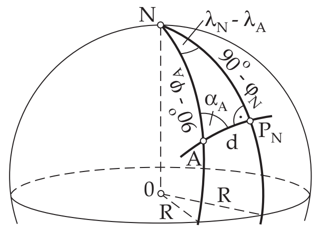


Figure 3.106

Remark: In certain cases an angle reduction is needed according to (3.205).

3. Arclength If the orthodrome goes through the points $A(\lambda_A,\ \varphi_A)$ and $B(\lambda_B,\ \varphi_B)$, the cosine law for sides yields for the spherical distance d or for the arclength between the two points:

$$d = \arccos[\sin\varphi_A \sin\varphi_B + \cos\varphi_A \cos\varphi_B \cos(\lambda_B - \lambda_A)]. \tag{3.207a}$$

This central angle can be converted into a length considering the radius of the Earth R:

$$d = \arccos[\sin\varphi_A \sin\varphi_B + \cos\varphi_A \cos\varphi_B \cos(\lambda_B - \lambda_A)] \cdot \frac{\pi R}{180^\circ}. \tag{3.207b}$$

4. Course Angle Using the sine and cosine laws for sides to calculate $\sin\alpha_A$ and $\cos\alpha_A$, gives the final result for the course angle α_A after division:

$$\alpha_A = \arctan \frac{\cos\varphi_A \cos\varphi_B \sin(\lambda_B - \lambda_A)}{\sin\varphi_B - \sin\varphi_A \cos d}. \tag{3.208}$$

Remark: With the formulas (3.207a), (3.208), (3.204a) and (3.204b), the coordinates of the point P_N closest to the north pole can be calculated for the orthodrome given by the two points A and B.

5. Intersection Point with a Parallel Circle For the intersection points $X_1(\lambda_{X_1},\ \varphi_X)$ and $X_2(\lambda_{X_2},\ \varphi_X)$ of an orthodrome with the parallel circle $\varphi = \varphi_X$ one gets from (3.203):

$$\lambda_{X_\nu} = \lambda_\mathrm{N} \mp \arccos \frac{\tan\varphi_X}{\tan\varphi_\mathrm{N}} \quad (\nu = 1, 2). \tag{3.209}$$

From the Neper rule (3.4.3.2,**2.**, p. 170) used for the intersection angles α_{X_1} and α_{X_2}, by which an orthodrome with a point $P_\mathrm{N}(\lambda_\mathrm{N},\ \varphi_\mathrm{N})$ closest to the north pole intersects the parallel circle $\varphi = \varphi_X$:

$$|\alpha_{X_\nu}| = \arcsin \frac{\cos\varphi_\mathrm{N}}{\cos\varphi_X} \qquad (\nu = 1, 2). \tag{3.210}$$

The argument in the arc sine function must be extremal with respect to the variable φ_X for the minimal course angle $|\alpha_{\min}|$. This gives: $\sin\varphi_X = 0 \ \Rightarrow\ \varphi_X = 0$, i.e., the absolute value of the course angle is minimal at the intersection points of the equator:

$$|\alpha_{X_{\min}}| = 90^\circ - \varphi_\mathrm{N}. \tag{3.211}$$

Remark 1: Solutions of (3.209) exist only for $|\varphi_X| \le \varphi_\mathrm{N}$.

Remark 2: In certain cases a reduction of the angle is needed according to (3.205).

6. Intersection Point with a Meridian For the intersection point $Y(\lambda_Y,\ \varphi_Y)$ of an orthodrome with the meridian $\lambda = \lambda_Y$ according to (3.203) is given by

$$\varphi_Y = \arctan[\tan\varphi_\mathrm{N} \cos(\lambda_Y - \lambda_\mathrm{N})]. \tag{3.212}$$

2. Small Circle

1. Notion The definition of a small circle on the surface of the sphere must be discussed here in more detail than in 3.4.1.1, p. 160: The *small circle* is the locus of the points of the surface (of the sphere) at a spherical distance r $(r < 90^\circ)$ from a point fixed on the surface $M(\lambda_M, \varphi_M)$ **(Fig. 3.107)**. The

spherical center is denoted by M; r is called the *spherical radius of the small circle.*

The plane of the small circle is the base of a spherical cap with an altitude h (see 3.3.4, p. 158). The spherical center M is above the center of the small circle in the plane of the small circle. In the plane the circle has the *planar radius of the small circle* r_0 **(Fig. 3.108)**. Hence, parallel circles are special small circles with $\varphi_M = \pm 90°$.

■ For $r \to 90°$ the small circle tends to an orthodrome.

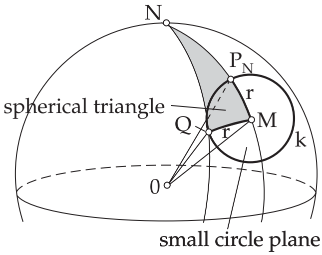


Figure 3.107

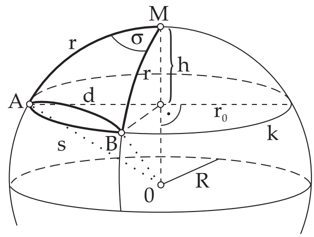


Figure 3.108

2. Equations of Small Circles As defining parameters either M and r can be used or the small circle point $P_{\mathrm{N}}(\lambda_{\mathrm{N}},\ \varphi_{\mathrm{N}})$ nearest the north pole and r. If the running point on the small circle is $Q(\lambda, \varphi)$, one gets the equation of the small circle with the cosine law for sides corresponding to **Fig. 3.107**:

$$\cos r = \sin\varphi \sin\varphi_M + \cos\varphi \cos\varphi_M \cos(\lambda - \lambda_M). \tag{3.213a}$$

From here, because of $\varphi_M = \varphi_{\mathrm{N}} - r$ and $\lambda_M = \lambda_{\mathrm{N}}$,

$$\cos r = \sin\varphi \sin(\varphi_{\mathrm{N}} - r) + \cos\varphi \cos(\varphi_{\mathrm{N}} - r) \cos(\lambda - \lambda_{\mathrm{N}}). \tag{3.213b}$$

■ **A:** For $\varphi_M = 90°$ the parallel circles are obtained from (3.213a), since $\cos r = \sin\varphi \Rightarrow \sin(90° - r) = \sin\varphi \Rightarrow \varphi = \text{const.}$

■ **B:** For $r \to 90°$ follow orthodromes from (3.213b).

3. Arclength The arclength s between two points $A(\lambda_A, \varphi_A)$ and $B(\lambda_B, \varphi_B)$ on a small circle k can be calculated corresponding to **Fig. 3.108** from the equalities $\dfrac{s}{\sigma} = \dfrac{2\pi r_0}{360°}$, $\cos d = \cos^2 r + \sin^2 r \cos\sigma$ and $r_0 = R \sin r$:

$$s = \sin r \arccos \frac{\cos d - \cos^2 r}{\sin^2 r} \cdot \frac{\pi R}{180°}. \tag{3.214}$$

■ For $r \to 90°$ the small circle becomes an orthodrome, and from (3.214) and (3.207b) it follows that $s = d$.

4. Course Angle According to **Fig. 3.109** the orthodrome passing through $A(\lambda_A, \varphi_A)$ and $M(\lambda_M, \varphi_M)$ intersects perpendicularly the small circle with radius r. For the course angle α_{Orth} of the orthodrome after (3.208)

$$\alpha_{\mathrm{Orth}} = \arctan \frac{\cos\varphi_A \cos\varphi_M \sin(\lambda_M - \lambda_A)}{\sin\varphi_M - \sin\varphi_A \cos r} \tag{3.215a}$$

is valid. So, one gets for the required course angle α_A of the small circle at the point A:

$$\alpha_A = (|\alpha_{\mathrm{Orth}}| - 90°)\,\mathrm{sign}(\alpha_{\mathrm{Orth}}). \tag{3.215b}$$

5. Intersection Points with a Parallel Circle For the geographical longitude of the intersection points $X_1(\lambda_{X_1},\ \varphi_X)$ and $X_2(\lambda_{X_2},\ \varphi_X)$ of the small circle with the parallel circle $\varphi = \varphi_X$ follows from

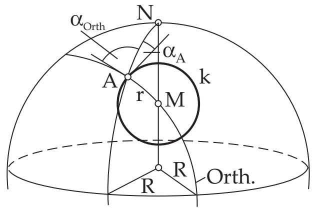


Figure 3.109

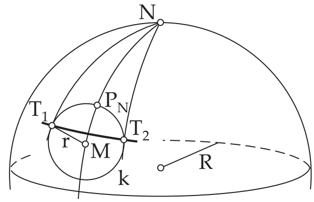


Figure 3.110

(3.213a):

$$\lambda_{X_\nu} = \lambda_M \mp \arccos \frac{\cos r - \sin\varphi_X \sin\varphi_M}{\cos\varphi_X \cos\varphi_M} \qquad (\nu = 1, 2). \tag{3.216}$$

Remark: In some cases an angle reduction is needed according to (3.205).

6. Tangential Point The small circle is touched by two meridians, the *tangential meridians*, at the *tangential points* $T_1(\lambda_{T_1},\ \varphi_T)$ and $T_2(\lambda_{T_2},\ \varphi_T)$ **(Fig. 3.110)**. Because for them the argument of the arc cosine in (3.216) must be extremal with respect to the variable φ_X, holds:

$$\varphi_T = \arcsin \frac{\sin\varphi_M}{\cos r}, \quad (3.217a) \qquad \lambda_{T_\nu} = \lambda_M \mp \arccos \frac{\cos r - \sin\varphi_X \sin\varphi_M}{\cos\varphi_X \cos\varphi_M} \quad (\nu = 1, 2). \quad (3.217b)$$

Remark: In certain cases an angle reduction is needed according to (3.205).

7. Intersection Points with a Meridian The calculation of the geographical latitudes of the intersection points $Y_1(\lambda_Y,\ \varphi_{Y_1})$ and $Y_2(\lambda_Y,\ \varphi_{Y_2})$ of the small circle with meridian $\lambda = \lambda_Y$ can be done according to (3.213a) with the equations

$$\varphi_{Y_\nu} = \arcsin \frac{-AC \pm B\sqrt{A^2 + B^2 - C^2}}{A^2 + B^2} \qquad (\nu = 1, 2), \tag{3.218a}$$

where the following notations are used:

$$A = \sin\varphi_M, \qquad B = \cos\varphi_M \cos(\lambda_Y - \lambda_M), \qquad C = -\cos r. \tag{3.218b}$$

For $A^2 + B^2 > C^2$, in general, there are two different solutions, from which one is missing if a pole is on the small circle.

If $A^2 + B^2 = C^2$ holds and there is no pole on the small circle, then the meridian touches the small circle at a tangential point with geographical latitude $\varphi_{Y_1} = \varphi_{Y_2} = \varphi_T$.

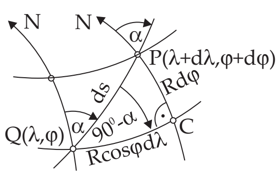


Figure 3.111

3. Loxodrome

1. Notion A spherical curve, intersecting all meridians with the same course angle, is called a *loxodrome* or *spherical helix*. So, parallels ($\alpha = 90°$) and meridians ($\alpha = 0°$) are special loxodromes.

2. Equation of the Loxodrome Fig. 3.111 shows a loxodrome with course angle α through the running point $Q(\lambda, \varphi)$ and the infinitesimally close point $P(\lambda + d\lambda, \varphi + d\varphi)$. The right-angled spherical triangle QCP can be considered as a plane triangle because of its small size. Then:

$$\tan\alpha = \frac{R\cos\varphi\, d\lambda}{R\, d\varphi} \ \Rightarrow\ d\lambda = \frac{\tan\alpha\, d\varphi}{\cos\varphi}. \tag{3.219a}$$

Considering that the loxodrome must go through the point $A(\lambda_A, \varphi_A)$, therefore, the equation of the loxodrome follows by integration:

$$\lambda - \lambda_A = \tan\alpha \ln \frac{\tan\left(45° + \frac{\varphi}{2}\right)}{\tan\left(45° + \frac{\varphi_A}{2}\right)} \cdot \frac{180°}{\pi} \qquad (\alpha \neq 90°). \tag{3.219b}$$

In particular if A is the intersection point $P_{\mathrm{E}}(\lambda_{\mathrm{E}}, 0°)$ of the loxodrome with the equator, then:

$$\lambda - \lambda_{\mathrm{E}} = \tan\alpha \ln\tan\left(45° + \frac{\varphi}{2}\right) \cdot \frac{180°}{\pi} \qquad (\alpha \neq 90°)\,. \tag{3.219c}$$

Remark: The calculation of λ_{E} can be done with (3.224).

3. Arclength From **Fig. 3.111** one can find the differential relation

$$\cos\alpha = \frac{R\,d\varphi}{ds} \quad\Rightarrow\quad ds = \frac{R\,d\varphi}{\cos\alpha}\,. \tag{3.220a}$$

Integration with respect to φ results in the arclength s of the arc segment with the endpoints $A(\lambda_A, \varphi_A)$ and $B(\lambda_B, \varphi_B)$:

$$s = \frac{|\varphi_B - \varphi_A|}{\cos\alpha} \cdot \frac{\pi R}{180°} \qquad (\alpha \neq 90°)\,. \tag{3.220b}$$

If A is the starting point and B is the endpoint, then from the given values A, α and s can be determined step-by-step first φ_B from (3.220b), then λ_B from (3.219b).

Approximation Formulas: According to **Fig. 3.111**, with $Q = A$ and $P = B$ one can get an approximation for the arclength l with the arithmetical mean of the geographical latitudes with (3.221a) and (3.221b):

$$\sin\alpha = \frac{\cos\frac{\varphi_A + \varphi_B}{2}(\lambda_B - \lambda_A)}{l} \cdot \frac{\pi R}{180°}\,. \tag{3.221a}$$

$$l = \frac{\cos\frac{\varphi_A + \varphi_B}{2}}{\sin\alpha}(\lambda_B - \lambda_A) \cdot \frac{\pi R}{180°}\,. \tag{3.221b}$$

4. Course Angle For the course angle α of the loxodrome through the points $A(\lambda_A, \varphi_A)$ and $B(\lambda_B,\varphi_B)$, or through $A(\lambda_A, \varphi_A)$ and its equator intersection point $P_{\mathrm{E}}(\lambda_{\mathrm{E}}, 0°)$ according to (3.219b) and (3.219c) the following holds:

$$\alpha = \arctan \frac{(\lambda_B - \lambda_A)}{\ln \frac{\tan\left(45° + \frac{\varphi_B}{2}\right)}{\tan\left(45° + \frac{\varphi_A}{2}\right)}} \cdot \frac{\pi}{180°}\,, \tag{3.222a}$$

$$\alpha = \arctan \frac{(\lambda_A - \lambda_{\mathrm{E}})}{\ln\tan\left(45° + \frac{\varphi_A}{2}\right)} \cdot \frac{\pi}{180°}\,. \tag{3.222b}$$

5. Intersection Point with a Parallel Circle Suppose a loxodrome passes through the point $A(\lambda_A, \varphi_A)$ with a course angle α. The intersection point $X(\lambda_X, \varphi_X)$ of the loxodrome with a parallel circle $\varphi = \varphi_X$ is calculated from (3.219b):

$$\lambda_X = \lambda_A + \tan\alpha \cdot \ln \frac{\tan\left(45° + \frac{\varphi_X}{2}\right)}{\tan\left(45° + \frac{\varphi_A}{2}\right)} \cdot \frac{180°}{\pi} \qquad (\alpha \neq 90°). \tag{3.223}$$

With (3.223) the intersection point with the equator $P_{\mathrm{E}}(\lambda_{\mathrm{E}}, 0°)$ is calculated as

$$\lambda_{\mathrm{E}} = \lambda_A - \tan\alpha \cdot \ln\tan\left(45° + \frac{\varphi_A}{2}\right) \cdot \frac{180°}{\pi} \qquad (\alpha \neq 90°)\,. \tag{3.224}$$

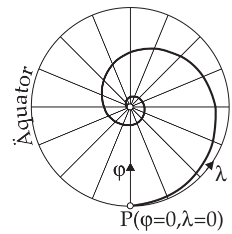


Figure 3.112

Remark: In certain cases an angle reduction is needed according to (3.205).

6. Intersection Point with a Meridian Loxodromes – except parallel circles and meridians – wind in a spiral form around the pole **(Fig. 3.112)**. The infinitely many intersection points $Y_\nu(\lambda_Y, \varphi_{Y_\nu})$ ($\nu \in \mathbb{Z}$) of the loxodrome passing through $A(\lambda_A, \varphi_A)$ with course angle α with the meridian $\lambda = \lambda_Y$ can be calculated from (3.219b):

$$\varphi_{Y_\nu} = 2 \arctan \left\{ \exp \left[\frac{\lambda_Y - \lambda_A + \nu \cdot 360°}{\tan \alpha} \cdot \frac{\pi}{180°} \right] \cdot \tan \left(45° + \frac{\varphi_A}{2} \right) \right\} - 90° \qquad (\nu \in \mathbb{Z}). \tag{3.225}$$

If A is the equator intersection point $P_{\rm E}(\lambda_{\rm E}, 0°)$ of the loxodrome, then simply holds

$$\varphi_{Y_\nu} = 2 \text{ arctan exp} \left[\frac{\lambda_Y - \lambda_{\rm E} + \nu \cdot 360°}{\tan \alpha} \cdot \frac{\pi}{180°} \right] - 90° \qquad (\nu \in \mathbb{Z}). \tag{3.226}$$

4. Intersection Points of Spherical Curves

1. Intersection Points of Two Orthodromes Suppose the considered orthodromes have points $P_{\rm N_1}(\lambda_{\rm N_1}, \varphi_{\rm N_1})$ and $P_{\rm N_2}(\lambda_{\rm N_2}, \varphi_{\rm N_2})$ closest to the north pole, where $P_{\rm N_1} \neq P_{\rm N_2}$ holds. Substituting the intersection point $S(\lambda_S, \varphi_S)$ in both orthodrome equations gives the system of equations:

$$\tan \varphi_{\rm N_1} \cos(\lambda_S - \lambda_{\rm N_1}) = \tan \varphi_S \,, \quad (3.227a) \qquad \tan \varphi_{\rm N_2} \cos(\lambda_S - \lambda_{\rm N_2}) = \tan \varphi_S \,. \quad (3.227b)$$

Elimination of φ_S and using the addition law for the cosine function yields:

$$\tan \lambda_S = -\frac{\tan \varphi_{\rm N_1} \cos \lambda_{\rm N_1} - \tan \varphi_{\rm N_2} \cos \lambda_{\rm N_2}}{\tan \varphi_{\rm N_1} \sin \lambda_{\rm N_1} - \tan \varphi_{\rm N_2} \sin \lambda_{\rm N_2}}. \tag{3.228}$$

The equation (3.228) has two solutions λ_{S_1} and λ_{S_2} in the domain $-180° < \lambda \leq 180°$ of the geographical longitude. The corresponding geographical latitudes can be got from (3.227a):

$$\varphi_{S_\nu} = \arctan[\tan \varphi_{\rm N_1} \cos(\lambda_{S_\nu} - \lambda_{\rm N_1})] \qquad (\nu = 1, 2). \tag{3.229}$$

The intersection points S_1 and S_2 are *antipodal points*, i.e., they are the mirror images of each other with respect to the centre of the sphere.

2. Intersection Points of Two Loxodromes Suppose the considered loxodromes have equator intersection points $P_{\rm E_1}(\lambda_{\rm E_1}, 0°)$ and $P_{\rm E_2}(\lambda_{\rm E_2}, 0°)$ and the course angles α_1 und α_2 ($\alpha_1 \neq \alpha_2$). Substituting the intersection point $S(\lambda_S, \varphi_S)$ in both loxodrome equations gives the system of equations:

$$\lambda_S - \lambda_{\rm E_1} = \tan \alpha_1 \cdot \ln \tan \left(45° + \frac{\varphi_S}{2} \right) \cdot \frac{180°}{\pi} \qquad (\alpha_1 \neq 90°), \tag{3.230a}$$

$$\lambda_S - \lambda_{\rm E_2} = \tan \alpha_2 \cdot \ln \tan \left(45° + \frac{\varphi_S}{2} \right) \cdot \frac{180°}{\pi} \qquad (\alpha_2 \neq 90°). \tag{3.230b}$$

Elimination of λ_S and expressing φ_S gives an equation with infinitely many solutions:

$$\varphi_{S_\nu} = 2 \text{ arctan exp} \left[\frac{\lambda_{\rm E_1} - \lambda_{E_2} + \nu \cdot 360°}{\tan \alpha_2 - \tan \alpha_1} \cdot \frac{\pi}{180°} \right] - 90° \qquad (\nu \in \mathbb{Z}). \tag{3.231}$$

The corresponding geographical longitudes λ_{S_ν} can be found by substituting φ_{S_ν} in (3.230a):

$$\lambda_{S_\nu} = \lambda_{\rm E_1} + \tan \alpha_1 \ln \tan \left(45° + \frac{\varphi_{S_\nu}}{2} \right) \cdot \frac{180°}{\pi} \qquad (\alpha_1 \neq 90°), \quad (\nu \in \mathbb{Z}). \tag{3.232}$$

Remark: In certain cases an angle reduction is needed according to (3.205).

3.5 Vector Algebra and Analytical Geometry

3.5.1 Vector Algebra

3.5.1.1 Definition of Vectors

1. Scalars and Vectors

Quantities whose values are real numbers are called *scalars*. Examples are mass, temperature, energy, and work (for scalar invariant see 3.5.1.5, p. 185, 1., p. 214 and 4.3.5.2, p. 288).
Quantities which can be completely described by a magnitude and by a direction in space are called *vectors*. Examples are power, velocity, acceleration, angular velocity, angular acceleration, and electrical and magnetic force. We represent vectors by directed line segments in space.
In this book the vectors of three-dimensional Euclidean space are denoted by $\vec{\mathbf{a}}$, and in matrix theory by $\underline{\mathbf{a}}$ (see also 4.1.3, p. 271).

2. Polar and Axial Vectors

Polar vectors represent quantities with magnitude and direction in space, such as speed and acceleration; *axial vectors* represent quantities with magnitude, direction in space, and direction of rotation, such as angular velocity and angular acceleration. In notation they are distinguished by a polar or by an axial arrow **(Fig. 3.113)**. In mathematical discussion they are treated in the same way.

3. Magnitude or Absolute Value and Direction in Space

For the quantitative description of vectors $\vec{\mathbf{a}}$ or $\underline{\mathbf{a}}$, as line segments between the initial and endpoint A and B resp., are used the *magnitude*, i.e., the *absolute value* $|\vec{\mathbf{a}}|$, the length of the line segment, and the *direction in space*, which is given by a set of angles.

4. Equality of Vectors

Two vectors $\vec{\mathbf{a}}$ and $\vec{\mathbf{b}}$ are equal if their magnitudes are the same, and they have the same direction, i.e., if they are parallel and oriented identically.
Opposite and equal vectors are of the same magnitude, but oppositely directed:

$$\overrightarrow{AB} = \vec{\mathbf{a}}, \quad \overrightarrow{BA} = -\vec{\mathbf{a}} \quad \text{but} \quad |\overrightarrow{AB}| = |\overrightarrow{BA}|. \tag{3.233}$$

Axial vectors have opposite and equal directions of rotation in this case.

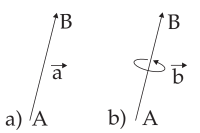


Figure 3.113

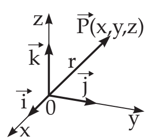


Figure 3.114

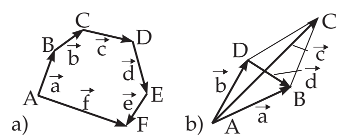


Figure 3.115

5. Free Vectors, Bound or Fixed Vectors, Sliding Vectors

A *free vector* is considered to be the same, i.e. its properties do not change, if it is translated parallel to itself, so its initial point can be an arbitrary point of the space. If the properties of a vector belong to a certain initial point, it is called a *bound or fixed vector*. A *sliding vector* can be translated only along the line it is already in.

6. Special Vectors

a) Unit Vector $\vec{\mathbf{a}}^0 = \vec{\mathbf{e}}$ is a vector with length or absolute value equal to 1. With it the vector $\vec{\mathbf{a}}$ can be expressed as a product of the magnitude and of a unit vector having the same direction as $\vec{\mathbf{a}}$:

$$\vec{\mathbf{a}} = \vec{\mathbf{e}}\,|\vec{\mathbf{a}}|. \tag{3.234}$$

The unit vectors $\vec{\mathbf{i}}, \vec{\mathbf{j}}, \vec{\mathbf{k}}$ or $\vec{\mathbf{e}}_i, \vec{\mathbf{e}}_j, \vec{\mathbf{e}}_k$ **(Fig. 3.114)** are often used to denote the three coordinate axes in the direction of increasing coordinate values.

In **Fig. 3.114** the directions given by the three unit vectors form an orthogonal *triple*. These unit vectors define an *orthogonal coordinate system* because for their scalar products

$$\vec{\mathbf{e_i}}\vec{\mathbf{e_j}} = \vec{\mathbf{e_i}}\vec{\mathbf{e_k}} = \vec{\mathbf{e_j}}\vec{\mathbf{e_k}} = 0 \tag{3.235}$$

is valid. Furthermore also

$$\vec{\mathbf{e_i}}\vec{\mathbf{e_i}} = \vec{\mathbf{e_j}}\vec{\mathbf{e_j}} = \vec{\mathbf{e_k}}\vec{\mathbf{e_k}} = 1, \tag{3.236}$$

holds, i.e. it is an *orthonormal coordinate system*. (For more about scalar product see (3.248).)

b) Null Vector $\vec{\mathbf{0}}$ or zero vector is the vector whose magnitude is equal to 0, i.e., its initial and endpoint coincide, and its direction is not defined.

c) Radius Vector $\vec{\mathbf{r}}$ or *position vector* of a point P is the vector $\overrightarrow{0P}$ with the initial point at the origin and endpoint at P **(Fig. 3.114)**. In this case the origin is also called a *pole* or *polar point*. The point P is defined uniquely by its radius vector.

d) Collinear Vectors are parallel to the same line.

e) Coplanar Vectors are parallel to the same plane. They satisfy the equality (3.260).

3.5.1.2 Calculation Rules for Vectors

1. Sum of Vectors

a) The Sum of Two Vectors $\overrightarrow{AB} = \vec{\mathbf{a}}$ and $\overrightarrow{AD} = \vec{\mathbf{b}}$ can be represented also as the diagonal of the parallelogram $ABCD$, as the vector $\overrightarrow{AC} = \vec{\mathbf{c}}$ in **Fig. 3.115b**. The most important properties of the sum of two vectors are the commutative law and the *triangle inequality*:

$$\vec{\mathbf{a}} + \vec{\mathbf{b}} = \vec{\mathbf{b}} + \vec{\mathbf{a}}, \quad |\,\vec{\mathbf{a}} + \vec{\mathbf{b}}\,| \le |\,\vec{\mathbf{a}}\,| + |\,\vec{\mathbf{b}}\,|. \tag{3.237a}$$

b) The Sum of Several Vectors $\vec{\mathbf{a}}, \vec{\mathbf{b}}, \vec{\mathbf{c}}, \ldots, \vec{\mathbf{e}}$ is the vector $\vec{\mathbf{f}} = \overrightarrow{AF}$, which closes the broken line composed of the vectors from $\vec{\mathbf{a}}$ to $\vec{\mathbf{e}}$ as in **Fig. 3.115a**. For n vectors $\vec{\mathbf{a}}_i$ $(i = 1, 2, \ldots, n)$ holds:

$$\sum_{i=1}^{n} \vec{\mathbf{a}}_i = \vec{\mathbf{f}}\,. \tag{3.237b}$$

Important properties of the sum of several vectors are the commutative law and the associative law of addition. For three vectors holds:

$$\vec{\mathbf{a}} + \vec{\mathbf{b}} + \vec{\mathbf{c}} = \vec{\mathbf{c}} + \vec{\mathbf{b}} + \vec{\mathbf{a}}, \quad (\vec{\mathbf{a}} + \vec{\mathbf{b}}) + \vec{\mathbf{c}} = \vec{\mathbf{a}} + (\vec{\mathbf{b}} + \vec{\mathbf{c}}). \tag{3.237c}$$

c) The Difference of Two Vectors $\vec{\mathbf{a}} - \vec{\mathbf{b}}$ can be considered as the sum of the vectors $\vec{\mathbf{a}}$ und $-\vec{\mathbf{b}}$, i.e.,

$$\vec{\mathbf{a}} - \vec{\mathbf{b}} = \vec{\mathbf{a}} + (-\vec{\mathbf{b}}) = \vec{\mathbf{d}} \tag{3.237d}$$

which is the other diagonal of the parallelogram **(Fig. 3.115b)**. The most important properties of the difference of two vectors are:

$$\vec{\mathbf{a}} - \vec{\mathbf{a}} = \vec{\mathbf{0}} \quad \text{(null vector)}, \qquad |\,\vec{\mathbf{a}} - \vec{\mathbf{b}}\,| \ge |\,|\,\vec{\mathbf{a}}\,| - |\,\vec{\mathbf{b}}\,|\,|. \tag{3.237e}$$

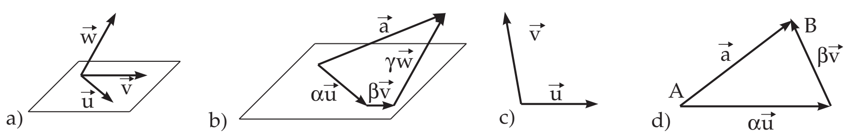


Figure 3.116

2. Multiplication of a Vector by a Scalar, Linear Combination

The products $\alpha\,\vec{\mathbf{a}}$ and $\vec{\mathbf{a}}\,\alpha$ are equal to each other and they are parallel (collinear) to $\vec{\mathbf{a}}$. The length (absolute value) of the product vector is equal to $|\alpha||\vec{\mathbf{a}}|$. For $\alpha > 0$ the product vector has the same

direction as $\vec{\mathbf{a}}$; for $\alpha < 0$ it has the opposite one. The most important properties of the product of vectors by scalars are:

$$\alpha\,\vec{\mathbf{a}} = \vec{\mathbf{a}}\,\alpha, \quad \alpha\,\beta\,\vec{\mathbf{a}} = \beta\,\alpha\,\vec{\mathbf{a}}, \quad (\alpha+\beta)\,\vec{\mathbf{a}} = \alpha\,\vec{\mathbf{a}} + \beta\,\vec{\mathbf{a}}, \quad \alpha\,(\vec{\mathbf{a}}+\vec{\mathbf{b}}) = \alpha\,\vec{\mathbf{a}} + \alpha\,\vec{\mathbf{b}}. \tag{3.238a}$$

The *linear combination* of the vectors $\vec{\mathbf{a}}, \vec{\mathbf{b}}, \vec{\mathbf{c}}, \dots, \vec{\mathbf{d}}$ with the scalars $\alpha, \beta, \dots, \delta$ is the vector

$$\vec{\mathbf{k}} = \alpha\,\vec{\mathbf{a}} + \beta\,\vec{\mathbf{b}} + \cdots + \delta\,\vec{\mathbf{d}}. \tag{3.238b}$$

3. Decomposition of Vectors

In three-dimensional space every vector $\vec{\mathbf{a}}$ can be decomposed uniquely into a sum of three vectors, which are parallel to the three given non-coplanar vectors $\vec{\mathbf{u}}, \vec{\mathbf{v}}, \vec{\mathbf{w}}$ **(Fig. 3.116a,b)**:

$$\vec{\mathbf{a}} = \alpha\,\vec{\mathbf{u}} + \beta\,\vec{\mathbf{v}} + \gamma\,\vec{\mathbf{w}}. \tag{3.239a}$$

The summands $\alpha\vec{\mathbf{u}}$, $\beta\vec{\mathbf{v}}$ and $\gamma\vec{\mathbf{w}}$ are called the *components* of the decomposition, the scalar factors α, β and γ are the *coefficients*. When all the vectors are parallel to a plane one can write

$$\vec{\mathbf{a}} = \alpha\,\vec{\mathbf{u}} + \beta\,\vec{\mathbf{v}} \tag{3.239b}$$

with two non-collinear vectors $\vec{\mathbf{u}}$ and $\vec{\mathbf{v}}$ being parallel to the same plane **(Fig. 3.116c,d)**.

3.5.1.3 Coordinates of a Vector

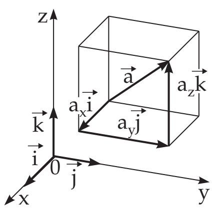


Figure 3.117

1. Cartesian Coordinates According to (3.239a) every vector $\overrightarrow{AB} = \vec{\mathbf{a}}$ can be decomposed uniquely into a sum of vectors parallel to the basis vectors of the coordinate system $\vec{\mathbf{i}}, \vec{\mathbf{j}}, \vec{\mathbf{k}}$ or $\vec{\mathbf{e}}_i, \vec{\mathbf{e}}_j, \vec{\mathbf{e}}_k$:

$$\vec{\mathbf{a}} = a_x\vec{\mathbf{i}} + a_y\vec{\mathbf{j}} + a_z\vec{\mathbf{k}} = a_x\vec{\mathbf{e}}_i + a_y\vec{\mathbf{e}}_j + a_z\vec{\mathbf{e}}_k, \tag{3.240a}$$

where the scalars a_x, a_y and a_z are the *Cartesian coordinates* of the vector $\vec{\mathbf{a}}$ in the system with the unit vectors $\vec{\mathbf{e}}_i$, $\vec{\mathbf{e}}_j$ and $\vec{\mathbf{e}}_k$. Also is written

$$\vec{\mathbf{a}} = \{a_x, a_y, a_z\} \quad \text{or} \quad \vec{\mathbf{a}}(a_x, a_y, a_z). \tag{3.240b}$$

The three directions defined by the unit vectors form an orthogonal *direction triple*. The components of a vector are the projections of this vector on the coordinate axes **(Fig. 3.117)**.

The coordinates of a linear combination of several vectors are the same linear combination of the coordinates of these vectors, so the vector equation (3.238b) corresponds to the following coordinate equations:

$$\begin{aligned} k_x &= \alpha\,a_x + \beta\,b_x + \cdots + \delta\,d_x, \\ k_y &= \alpha\,a_y + \beta\,b_y + \cdots + \delta\,d_y, \\ k_z &= \alpha\,a_z + \beta\,b_z + \cdots + \delta\,d_z. \end{aligned} \tag{3.241}$$

For the coordinates of the sum and of the difference of two vectors

$$\vec{\mathbf{c}} = \vec{\mathbf{a}} \pm \vec{\mathbf{b}} \tag{3.242a}$$

the equalities

$$c_x = a_x \pm b_x, \quad c_y = a_y \pm b_y, \quad c_z = a_z \pm a_z \tag{3.242b}$$

are valid. The radius vector $\vec{\mathbf{r}}$ of the point $P(x, y, z)$ has the Cartesian coordinates of this point:

$$r_x = x, \quad r_y = y, \quad r_z = z; \quad \vec{\mathbf{r}} = x\,\vec{\mathbf{i}} + y\,\vec{\mathbf{j}} + z\,\vec{\mathbf{k}}. \tag{3.243}$$

2. Affine Coordinates are a generalization of Cartesian coordinates with respect to a system of linearly independent but not necessarily orthogonal vectors, i.e., to three non-coplanar basis vectors $\vec{\mathbf{e}}_1, \vec{\mathbf{e}}_2, \vec{\mathbf{e}}_3$. The coefficients are a^1, a^2, a^3, where the upper indices are not exponents. Similarly to (3.240a,b) for $\vec{\mathbf{a}}$ holds

$$\vec{\mathbf{a}} = a^1\,\vec{\mathbf{e}}_1 + a^2\,\vec{\mathbf{e}}_2 + a^3\,\vec{\mathbf{e}}_3 \quad (3.244a) \qquad \text{or} \quad \vec{\mathbf{a}} = \left\{a^1,\, a^2, a^3\right\} \text{ or } \vec{\mathbf{a}}\left(a^1,\, a^2, a^3\right). \tag{3.244b}$$

This notation is especially suitable as the scalars a^1, a^2, a^3 are the contravariant coordinates of a vector

(see 3.5.1.8, p. 188). For $\vec{\mathbf{e}}_1 = \vec{\mathbf{i}}$, $\vec{\mathbf{e}}_2 = \vec{\mathbf{j}}$, $\vec{\mathbf{e}}_3 = \vec{\mathbf{k}}$ the formulas (3.244a,b) become (3.240a,c). For the linear combination of vectors (3.238b) just as for the sum and difference of two vectors (3.242a,b) in analogy to (3.241) the same coordinate equations are valid:

$$k^1 = \alpha\, a^1 + \beta\, b^1 + \cdots + \delta\, d^1,$$
$$k^2 = \alpha\, a^2 + \beta\, b^2 + \cdots + \delta\, d^2, \qquad (3.245)$$
$$k^3 = \alpha\, a^3 + \beta\, b^3 + \cdots + \delta\, d^3;$$
$$c^1 = a^1 \pm b^1, \quad c^2 = a^2 \pm b^2, \quad c^3 = a^3 \pm b^3. \qquad (3.246)$$

3.5.1.4 Directional Coefficient

The *directional coefficient* of a vector $\vec{\mathbf{a}}$ along a vector $\vec{\mathbf{b}}$ is the scalar product

$$a_b = \vec{\mathbf{a}}\,\vec{\mathbf{b}}^0 = |\vec{\mathbf{a}}|\,\cos\varphi, \qquad (3.247)$$

where $\vec{\mathbf{b}}^0 = \dfrac{b}{|\vec{\mathbf{b}}|}$ is the unit vector in the direction of $\vec{\mathbf{b}}$ and φ is the angle between $\vec{\mathbf{a}}$ and $\vec{\mathbf{b}}$.

The directional coefficient represents the projection of $\vec{\mathbf{a}}$ on $\vec{\mathbf{b}}$.

■ In the Cartesian coordinate system the directional coefficients of the vector $\vec{\mathbf{a}}$ along the x, y, z axes are the coordinates a_x, a_y, a_z. This statement is usually not true in a non-orthonormal coordinate system.

3.5.1.5 Scalar Product and Vector Product

1. Scalar product

The *scalar product* or *dot product* of two vectors $\vec{\mathbf{a}}$ and $\vec{\mathbf{b}}$ is defined by the equation

$$\vec{\mathbf{a}} \cdot \vec{\mathbf{b}} = \vec{\mathbf{a}}\,\vec{\mathbf{b}} = (\vec{\mathbf{a}}\,\vec{\mathbf{b}}) = |\vec{\mathbf{a}}|\,|\vec{\mathbf{b}}|\cos\varphi, \qquad (3.248)$$

where φ is the angle between $\vec{\mathbf{a}}$ and $\vec{\mathbf{b}}$ considering them with a common initial point **(Fig. 3.118)**. The value of a scalar product is a scalar.

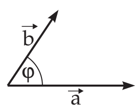


Figure 3.118

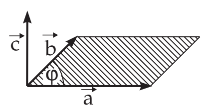


Figure 3.119

2. Vector Product

or *cross product* of the two vectors $\vec{\mathbf{a}}$ and $\vec{\mathbf{b}}$ is a vector $\vec{\mathbf{c}}$ such that it is perpendicular to the vectors $\vec{\mathbf{a}}$ and $\vec{\mathbf{b}}$, and in the order $\vec{\mathbf{a}}$, $\vec{\mathbf{b}}$, and $\vec{\mathbf{c}}$ the vectors form a right-hand system **(Fig. 3.119)**: If the vectors have the same initial point, then looking at the plane of $\vec{\mathbf{a}}$ and $\vec{\mathbf{b}}$ from the endpoint of $\vec{\mathbf{c}}$, the shortest rotation of $\vec{\mathbf{a}}$ in the direction of $\vec{\mathbf{b}}$ is counterclockwise. The vectors $\vec{\mathbf{a}}$, $\vec{\mathbf{b}}$, and $\vec{\mathbf{c}}$ have the same arrangement as the thumb, the forefinger, and the middle finger of the right hand. Therefore this is called the *right-hand rule*. The vector product (3.249a) has the magnitude (3.249b)

$$\vec{\mathbf{a}} \times \vec{\mathbf{b}} = [\vec{\mathbf{a}}\,\vec{\mathbf{b}}] = \vec{\mathbf{c}}, \qquad (3.249a) \qquad\qquad |\vec{\mathbf{c}}| = |\vec{\mathbf{a}}|\,|\vec{\mathbf{b}}|\,\sin\varphi, \qquad (3.249b)$$

where φ is the angle between $\vec{\mathbf{a}}$ and $\vec{\mathbf{b}}$. Numerically the length of $\vec{\mathbf{c}}$ is equal to the area of the parallelogram defined by the vectors $\vec{\mathbf{a}}$ and $\vec{\mathbf{b}}$.

3. Properties of the Products of Vectors

a) The Scalar Product is commutative:

$$\vec{\mathbf{a}}\,\vec{\mathbf{b}} = \vec{\mathbf{b}}\,\vec{\mathbf{a}}. \qquad (3.250)$$

b) The Vector Product is anti commutative (changes its sign by interchanging the factors):

$$\vec{\mathbf{a}} \times \vec{\mathbf{b}} = -(\vec{\mathbf{b}} \times \vec{\mathbf{a}}). \qquad (3.251)$$

c) Multiplication by a Scalar Scalars can be factored out:

$$\alpha(\vec{\mathbf{a}}\,\vec{\mathbf{b}}\,) = (\alpha\,\vec{\mathbf{a}})\,\vec{\mathbf{b}}, \qquad (3.252a) \qquad\qquad \alpha(\vec{\mathbf{a}}\times\vec{\mathbf{b}}\,) = (\alpha\,\vec{\mathbf{a}})\times\vec{\mathbf{b}}. \qquad (3.252b)$$

d) Associativity The scalar and vector products are not associative:

$$\vec{\mathbf{a}}\,(\vec{\mathbf{b}}\,\vec{\mathbf{c}}\,) \neq (\vec{\mathbf{a}}\,\vec{\mathbf{b}}\,)\,\vec{\mathbf{c}}, \qquad (3.253a) \qquad\qquad \vec{\mathbf{a}}\times(\vec{\mathbf{b}}\times\vec{\mathbf{c}}) \neq (\vec{\mathbf{a}}\times\vec{\mathbf{b}}\,)\times\vec{\mathbf{c}}. \qquad (3.253b)$$

e) Distributivity The scalar and vector products are distributive over addition:

$$\vec{\mathbf{a}}\,(\vec{\mathbf{b}}+\vec{\mathbf{c}}\,) = \vec{\mathbf{a}}\,\vec{\mathbf{b}}+\vec{\mathbf{a}}\,\vec{\mathbf{c}}, \qquad (3.254a)$$

$$\vec{\mathbf{a}}\times(\vec{\mathbf{b}}+\vec{\mathbf{c}}\,) = \vec{\mathbf{a}}\times\vec{\mathbf{b}}+\vec{\mathbf{a}}\times\vec{\mathbf{c}} \quad \text{and} \quad (\vec{\mathbf{b}}+\vec{\mathbf{c}}\,)\times\vec{\mathbf{a}} = \vec{\mathbf{b}}\times\vec{\mathbf{a}}+\vec{\mathbf{c}}\times\vec{\mathbf{a}}. \qquad (3.254b)$$

f) Orthogonality of Two Vectors Two vectors are perpendicular to each other ($\vec{\mathbf{a}}\perp\vec{\mathbf{b}}$) if the equality

$$\vec{\mathbf{a}}\,\vec{\mathbf{b}} = 0 \quad \text{holds, and neither } \vec{\mathbf{a}} \text{ nor } \vec{\mathbf{b}} \text{ are null vectors.} \qquad (3.255)$$

g) Collinearity of Two Vectors Two vectors are collinear ($\vec{\mathbf{a}}\,\|\,\vec{\mathbf{b}}$) if the equality

$$\vec{\mathbf{a}}\times\vec{\mathbf{b}} = \vec{\mathbf{0}} \qquad \text{holds, and neither } \vec{\mathbf{a}} \text{ nor } \vec{\mathbf{b}} \text{ are null vectors.} \qquad (3.256)$$

h) Multiplication of the same vectors:

$$\vec{\mathbf{a}}\,\vec{\mathbf{a}} = \vec{\mathbf{a}}^2 = a^2, \quad \vec{\mathbf{a}}\times\vec{\mathbf{a}} = \vec{\mathbf{0}}. \qquad (3.257)$$

i) Linear Combinations of Vectors can be multiplied in the same way as scalar polynomials (because of the distributive property), only one must be careful with the vector product. If interchanging the factors, also the signs are to be changed.

■ **A:** $(3\vec{\mathbf{a}}+5\vec{\mathbf{b}}-2\vec{\mathbf{c}})(\vec{\mathbf{a}}-2\vec{\mathbf{b}}-4\vec{\mathbf{c}}) = 3\vec{\mathbf{a}}^2+5\vec{\mathbf{b}}\vec{\mathbf{a}}-2\vec{\mathbf{c}}\vec{\mathbf{a}}-6\vec{\mathbf{a}}\vec{\mathbf{b}}-10\vec{\mathbf{b}}^2+4\vec{\mathbf{c}}\vec{\mathbf{b}}-12\vec{\mathbf{a}}\vec{\mathbf{c}}-20\vec{\mathbf{b}}\vec{\mathbf{c}}+8\vec{\mathbf{c}}^2$
$= 3\vec{\mathbf{a}}^2-10\vec{\mathbf{b}}^2+8\vec{\mathbf{c}}^2-\vec{\mathbf{a}}\vec{\mathbf{b}}-14\vec{\mathbf{a}}\vec{\mathbf{c}}-16\vec{\mathbf{b}}\vec{\mathbf{c}}.$

■ **B:** $(3\vec{\mathbf{a}}+5\vec{\mathbf{b}}-2\vec{\mathbf{c}})\times(\vec{\mathbf{a}}-2\vec{\mathbf{b}}-4\vec{\mathbf{c}}) = 3\vec{\mathbf{a}}\times\vec{\mathbf{a}}+5\vec{\mathbf{b}}\times\vec{\mathbf{a}}-2\vec{\mathbf{c}}\times\vec{\mathbf{a}}-6\vec{\mathbf{a}}\times\vec{\mathbf{b}}-10\vec{\mathbf{b}}\times\vec{\mathbf{b}}$
$+\,4\vec{\mathbf{c}}\times\vec{\mathbf{b}}-12\vec{\mathbf{a}}\times\vec{\mathbf{c}}-20\vec{\mathbf{b}}\times\vec{\mathbf{c}}+8\vec{\mathbf{c}}\times\vec{\mathbf{c}} = 0-5\vec{\mathbf{a}}\times\vec{\mathbf{b}}+2\vec{\mathbf{a}}\times\vec{\mathbf{c}}-6\vec{\mathbf{a}}\times\vec{\mathbf{b}}+0-4\vec{\mathbf{b}}\times\vec{\mathbf{c}}$
$-\,12\vec{\mathbf{a}}\times\vec{\mathbf{c}}-20\vec{\mathbf{b}}\times\vec{\mathbf{c}}+0 = -11\vec{\mathbf{a}}\times\vec{\mathbf{b}}-10\vec{\mathbf{a}}\times\vec{\mathbf{c}}-24\vec{\mathbf{b}}\times\vec{\mathbf{c}} = 11\vec{\mathbf{b}}\times\vec{\mathbf{a}}+10\vec{\mathbf{c}}\times\vec{\mathbf{a}}+24\vec{\mathbf{c}}\times\vec{\mathbf{b}}.$

j) Scalar Invariant is a scalar quantity if it does not change its value under a translation or a rotation of the coordinate system. The scalar product of two vectors is a scalar invariant.

■ **A:** The coordinates of a vector $\vec{\mathbf{a}} = \{a_1, a_2, a_3\}$ are not scalar invariants, because in different coordinate systems they can have different values.

■ **B:** The length of a vector $\vec{\mathbf{a}}$ is a scalar invariant, because it has the same value in different coordinate systems.

■ **C:** Since the scalar product of two vectors is a scalar invariant, the scalar product of a vector by itself is also a scalar invariant, i.e., $\vec{\mathbf{a}}\,\vec{\mathbf{a}} = |\vec{\mathbf{a}}|^2\cos\varphi = |\vec{\mathbf{a}}|^2$, because $\varphi = 0$.

3.5.1.6 Combination of Vector Products

1. Double Vector Product

The *double vector product* $\vec{\mathbf{a}}\times(\vec{\mathbf{b}}\times\vec{\mathbf{c}})$ results in a vector coplanar to $\vec{\mathbf{b}}$ and $\vec{\mathbf{c}}$:

$$\vec{\mathbf{a}}\times(\vec{\mathbf{b}}\times\vec{\mathbf{c}}) = \vec{\mathbf{b}}\,(\vec{\mathbf{a}}\,\vec{\mathbf{c}}) - \vec{\mathbf{c}}\,(\vec{\mathbf{a}}\,\vec{\mathbf{b}}\,). \qquad (3.258)$$

2. Mixed Product

The *mixed product* $(\vec{\mathbf{a}}\times\vec{\mathbf{b}})\,\vec{\mathbf{c}}$, which is also called the *triple product*, results in a scalar whose absolute value is numerically equal to the volume of the parallelepipedon defined by the three vectors; the result is positive if $\vec{\mathbf{a}}$, $\vec{\mathbf{b}}$, and $\vec{\mathbf{c}}$ form a right-hand system, negative otherwise. Parentheses and crosses can be

omitted:

$$(\vec{\mathbf{a}} \times \vec{\mathbf{b}})\vec{\mathbf{c}} = \vec{\mathbf{a}}(\vec{\mathbf{b}} \times \vec{\mathbf{c}}) = \vec{\mathbf{a}}\,\vec{\mathbf{b}}\,\vec{\mathbf{c}} = \vec{\mathbf{b}}\,\vec{\mathbf{c}}\,\vec{\mathbf{a}} = \vec{\mathbf{c}}\,\vec{\mathbf{a}}\vec{\mathbf{b}} = -\vec{\mathbf{a}}\vec{\mathbf{c}}\,\vec{\mathbf{b}} = -\vec{\mathbf{b}}\,\vec{\mathbf{a}}\vec{\mathbf{c}} = -\vec{\mathbf{c}}\,\vec{\mathbf{b}}\,\vec{\mathbf{a}}. \tag{3.259}$$

The interchange of any two terms results in a change of sign; the cyclic permutation of all three terms does not affect the result.
For *coplanar vectors*, i.e., if $\vec{\mathbf{a}}$ is parallel to the plane defined by $\vec{\mathbf{b}}$ and $\vec{\mathbf{c}}$, holds:

$$\vec{\mathbf{a}}\,(\vec{\mathbf{b}} \times \vec{\mathbf{c}}) = 0. \tag{3.260}$$

3. Formulas for Multiple Products

a) Lagrange Identity: $(\vec{\mathbf{a}} \times \vec{\mathbf{b}})(\vec{\mathbf{c}} \times \vec{\mathbf{d}}) = (\vec{\mathbf{a}}\,\vec{\mathbf{c}})\,(\vec{\mathbf{b}}\,\vec{\mathbf{d}}) - (\vec{\mathbf{b}}\,\vec{\mathbf{c}})\,(\vec{\mathbf{a}}\,\vec{\mathbf{d}}),$ (3.261)

b) $$\vec{\mathbf{a}}\,\vec{\mathbf{b}}\,\vec{\mathbf{c}} \cdot \vec{\mathbf{e}}\,\vec{\mathbf{f}}\,\vec{\mathbf{g}} = \begin{vmatrix} \vec{\mathbf{a}}\vec{\mathbf{e}} & \vec{\mathbf{a}}\vec{\mathbf{f}} & \vec{\mathbf{a}}\vec{\mathbf{g}} \\ \vec{\mathbf{b}}\vec{\mathbf{e}} & \vec{\mathbf{b}}\vec{\mathbf{f}} & \vec{\mathbf{b}}\vec{\mathbf{g}} \\ \vec{\mathbf{c}}\,\vec{\mathbf{e}} & \vec{\mathbf{c}}\vec{\mathbf{f}} & \vec{\mathbf{c}}\,\vec{\mathbf{g}} \end{vmatrix}. \tag{3.262}$$

4. Formulas for Products in Cartesian Coordinates

If the vectors $\vec{\mathbf{a}}$, $\vec{\mathbf{b}}$, $\vec{\mathbf{c}}$ are given by Cartesian coordinates as

$$\vec{\mathbf{a}} = \{a_x, a_y, a_z\}, \quad \vec{\mathbf{b}} = \{b_x, b_y, b_z\}, \quad \vec{\mathbf{c}} = \{c_x, c_y, c_z\}, \tag{3.263}$$

the calculation of the products can be made by the following formulas:

1. Scalar Product: $\vec{\mathbf{a}}\,\vec{\mathbf{b}} = a_x b_x + a_y b_y + a_z b_z.$ (3.264)

2. Vector Product:
$$\vec{\mathbf{a}} \times \vec{\mathbf{b}} = (a_y b_z - a_z b_y)\,\vec{\mathbf{i}} + (a_z b_x - a_x b_z)\,\vec{\mathbf{j}} + (a_x b_y - a_y b_x)\,\vec{\mathbf{k}} = \begin{vmatrix} \vec{\mathbf{i}} & \vec{\mathbf{j}} & \vec{\mathbf{k}} \\ a_x & a_y & a_z \\ b_x & b_y & b_z \end{vmatrix}. \tag{3.265}$$

3. Mixed Product: $$\vec{\mathbf{a}}\,\vec{\mathbf{b}}\,\vec{\mathbf{c}} = \begin{vmatrix} a_x & a_y & a_z \\ b_x & b_y & b_z \\ c_x & c_y & c_z \end{vmatrix}. \tag{3.266}$$

5. Formulas for Products in Affine Coordinates

1. Metric Coefficients and Reciprocal System of Vectors If the affine coordinates of two vectors $\vec{\mathbf{a}}$ and $\vec{\mathbf{b}}$ in the system of $\vec{\mathbf{e}}_1$, $\vec{\mathbf{e}}_2$, $\vec{\mathbf{e}}_3$ are given, i.e.,

$$\vec{\mathbf{a}} = a^1\,\vec{\mathbf{e}}_1 + a^2\,\vec{\mathbf{e}}_2 + a^3\,\vec{\mathbf{e}}_3, \quad \vec{\mathbf{b}} = b^1\,\vec{\mathbf{e}}_1 + b^2\,\vec{\mathbf{e}}_2 + b^3\,\vec{\mathbf{e}}_3, \tag{3.267}$$

and there is to calculate the scalar product

$$\vec{\mathbf{a}}\,\vec{\mathbf{b}} = a^1 b^1\,\vec{\mathbf{e}}_1\,\vec{\mathbf{e}}_1 + a^2\,b^2\,\vec{\mathbf{e}}_2\,\vec{\mathbf{e}}_2 + a^3\,b^3\,\vec{\mathbf{e}}_3\,\vec{\mathbf{e}}_3 + \left(a^1\,b^2 + a^2\,b^1\right)\vec{\mathbf{e}}_1\,\vec{\mathbf{e}}_2 + \left(a^2\,b^3 + a^3\,b^2\right)\vec{\mathbf{e}}_2\,\vec{\mathbf{e}}_3 + \left(a^3\,b^1 + a^1\,b^3\right)\vec{\mathbf{e}}_3\,\vec{\mathbf{e}}_1 \tag{3.268}$$

or the vector product

$$\vec{\mathbf{a}} \times \vec{\mathbf{b}} = \left(a^2\,b^3 - a^3\,b^2\right)\vec{\mathbf{e}}_2 \times \vec{\mathbf{e}}_3 + \left(a^3\,b^1 - a^1\,b^3\right)\vec{\mathbf{e}}_3 \times \vec{\mathbf{e}}_1 + \left(a^1\,b^2 - a^2\,b^1\right)\vec{\mathbf{e}}_1 \times \vec{\mathbf{e}}_2, \tag{3.269a}$$

with the equalities

$$\vec{\mathbf{e}}_1 \times \vec{\mathbf{e}}_1 = \vec{\mathbf{e}}_2 \times \vec{\mathbf{e}}_2 = \vec{\mathbf{e}}_3 \times \vec{\mathbf{e}}_3 = \vec{\mathbf{0}}, \tag{3.269b}$$

then the pairwise products of the coordinate vectors must be known. For the scalar product these are the six *metric coefficients* (numbers)

$$g_{11} = \vec{\mathbf{e}}_1\,\vec{\mathbf{e}}_1, \quad g_{22} = \vec{\mathbf{e}}_2\,\vec{\mathbf{e}}_2, \quad g_{33} = \vec{\mathbf{e}}_3\,\vec{\mathbf{e}}_3,$$
$$g_{12} = \vec{\mathbf{e}}_1\,\vec{\mathbf{e}}_2 = \vec{\mathbf{e}}_2\,\vec{\mathbf{e}}_1, \quad g_{23} = \vec{\mathbf{e}}_2\,\vec{\mathbf{e}}_3 = \vec{\mathbf{e}}_3\,\vec{\mathbf{e}}_2, \quad g_{31} = \vec{\mathbf{e}}_3\,\vec{\mathbf{e}}_1 = \vec{\mathbf{e}}_1\,\vec{\mathbf{e}}_3 \tag{3.270}$$

and for the vector product the three vectors

$$\vec{\mathbf{e}}^1 = \Omega\,(\vec{\mathbf{e}}_2 \times \vec{\mathbf{e}}_3)\,, \quad \vec{\mathbf{e}}^2 = \Omega\,(\vec{\mathbf{e}}_3 \times \vec{\mathbf{e}}_1)\,, \quad \vec{\mathbf{e}}^3 = \Omega\,(\vec{\mathbf{e}}_1 \times \vec{\mathbf{e}}_2)\,, \tag{3.271a}$$

which are the three *reciprocal vectors* with respect to $\vec{\mathbf{e}}_1$, $\vec{\mathbf{e}}_2$, $\vec{\mathbf{e}}_3$, where the coefficient

$$\Omega = \frac{1}{\vec{\mathbf{e}}_1\,\vec{\mathbf{e}}_2\,\vec{\mathbf{e}}_3}, \tag{3.271b}$$

is the reciprocal value of the mixed product of the coordinate vectors. This notation serves only as a shorter way of writing in the following discussion. Calculations with the coefficients will be easy to perform with the help of the **multiplication Tables 3.13** and **3.14** for the basis vectors .

Table 3.13 Scalar product of basis vectors

	$\vec{\mathbf{e}}_1$	$\vec{\mathbf{e}}_2$	$\vec{\mathbf{e}}_3$
$\vec{\mathbf{e}}_1$	g_{11}	g_{12}	g_{13}
$\vec{\mathbf{e}}_2$	g_{21}	g_{22}	g_{23}
$\vec{\mathbf{e}}_3$	g_{31}	g_{32}	g_{33}

$(g_{ki} = g_{ik})$

Table 3.14 Vector product of basis vectors

Multiplicands \ Multipliers	$\vec{\mathbf{e}}_1$	$\vec{\mathbf{e}}_2$	$\vec{\mathbf{e}}_3$
$\vec{\mathbf{e}}_1$	0	$\frac{\vec{\mathbf{e}}^3}{\Omega}$	$-\frac{\vec{\mathbf{e}}^2}{\Omega}$
$\vec{\mathbf{e}}_2$	$-\frac{\vec{\mathbf{e}}^3}{\Omega}$	0	$\frac{\vec{\mathbf{e}}^1}{\Omega}$
$\vec{\mathbf{e}}_3$	$\frac{\vec{\mathbf{e}}^2}{\Omega}$	$-\frac{\vec{\mathbf{e}}^1}{\Omega}$	0

Table 3.15 Scalar product of reciprocal basis vectors

	$\vec{\mathbf{i}}$	$\vec{\mathbf{j}}$	$\vec{\mathbf{k}}$
$\vec{\mathbf{i}}$	1	0	0
$\vec{\mathbf{j}}$	0	1	0
$\vec{\mathbf{k}}$	0	0	1

Table 3.16 Vector product of reciprocal basis vectors

Multiplicands \ Multipliers	$\vec{\mathbf{i}}$	$\vec{\mathbf{j}}$	$\vec{\mathbf{k}}$
$\vec{\mathbf{i}}$	0	$\vec{\mathbf{k}}$	$-\vec{\mathbf{j}}$
$\vec{\mathbf{j}}$	$-\vec{\mathbf{k}}$	0	$\vec{\mathbf{i}}$
$\vec{\mathbf{k}}$	$\vec{\mathbf{j}}$	$-\vec{\mathbf{i}}$	0

2. Application to Cartesian Coordinates The Cartesian coordinates are a special case of affine coordinates. From **Tables 3.15** and **3.16** for the basis vectors holds

$$\vec{\mathbf{e}}_1 = \vec{\mathbf{i}}, \quad \vec{\mathbf{e}}_2 = \vec{\mathbf{j}}, \quad \vec{\mathbf{e}}_3 = \vec{\mathbf{k}} \tag{3.272a}$$

with the metric coefficients

$$g_{11} = g_{22} = g_{33} = 1, \quad g_{12} = g_{23} = g_{31} = 0, \quad \Omega = \frac{1}{\vec{\mathbf{i}}\,\vec{\mathbf{j}}\,\vec{\mathbf{k}}} = 1, \tag{3.272b}$$

and the reciprocal basis vectors

$$\vec{\mathbf{e}}^1 = \vec{\mathbf{i}}, \quad \vec{\mathbf{e}}^2 = \vec{\mathbf{j}}, \quad \vec{\mathbf{e}}^3 = \vec{\mathbf{k}}. \tag{3.272c}$$

So the basis vectors coincide with the reciprocal basis vectors of the coordinate system, or, in other words, in the Cartesian coordinate system the basis vector system is its own reciprocal system.

3. Scalar Product of Vectors Given by Coordinates

$$\vec{\mathbf{a}}\vec{\mathbf{b}} = \sum_{m=1}^{3}\sum_{n=1}^{3} g_{mn}a^m\,b^n = g_{\alpha\beta}a^\alpha\,b^\beta. \tag{3.273}$$

For Cartesian coordinates (3.273) coincides with (3.264).
After the second equality in (3.273) a shorter notation for the sum was applied which is often used in tensor calculations (see 4.3.1, **2.**, p. 280): instead of the complete sum one writes only a characteristic term so that the sum should be calculated for repeated indices, i.e., for the indices appearing once down and once up. Sometimes the summation indices are denoted by Greek letters; here they have the values from 1 until 3. Consequently holds

$$g_{\alpha\beta}a^{\alpha}\,b^{\beta} = g_{11}a^1\,b^1 + g_{12}a^1\,b^2 + g_{13}a^1\,b^3 + g_{21}a^2b^1 + g_{22}a^2\,b^2 + g_{23}a^2\,b^3 + g_{31}a^3b^1 + g_{32}a^3\,b^2 + g_{33}a^3\,b^3. \tag{3.274}$$

4. Vector Product of Vectors Given by Coordinates In accordance with (3.269a)

$$\vec{\mathbf{a}} \times \vec{\mathbf{b}} = \vec{\mathbf{e}}_1\,\vec{\mathbf{e}}_2\,\vec{\mathbf{e}}_3 \begin{vmatrix} \vec{\mathbf{e}}^1 & \vec{\mathbf{e}}^2 & \vec{\mathbf{e}}^3 \\ a^1 & a^2 & a^3 \\ b^1 & b^2 & b^3 \end{vmatrix} = \vec{\mathbf{e}}_1\,\vec{\mathbf{e}}_2\,\vec{\mathbf{e}}_3 \left[(a^2\,b^3 - a^3\,b^2)\vec{\mathbf{e}}^1 + (a^3\,b^1 - a^1\,b^3)\,\vec{\mathbf{e}}^2 + (a^1\,b^2 - a^2\,b^1)\vec{\mathbf{e}}^3\right] \tag{3.275}$$

is valid. For Cartesian coordinates (3.275) coincides with (3.265).

5. Mixed Product of Vectors Given by Coordinates In accordance with (3.269a) holds

$$\vec{\mathbf{a}}\,\vec{\mathbf{b}}\,\vec{\mathbf{c}} = \vec{\mathbf{e}}_1\,\vec{\mathbf{e}}_2\,\vec{\mathbf{e}}_3 \begin{vmatrix} a^1 & a^2 & a^3 \\ b^1 & b^2 & b^3 \\ c^1 & c^2 & c^3 \end{vmatrix}. \tag{3.276}$$

For Cartesian coordinates (3.276) coincides with (3.266).

3.5.1.7 Vector Equations

Table 3.17 contains a summary of the simplest vector equations. In this table $\vec{\mathbf{a}}$, $\vec{\mathbf{b}}$, $\vec{\mathbf{c}}$ are given vectors, $\vec{\mathbf{x}}$ is the unknown vector, α, β, γ are given scalars, and x, y, z are the unknown scalars to be calculated.

3.5.1.8 Covariant and Contravariant Coordinates of a Vector

1. Definitions The affine coordinates a^1, a^2, a^3 of a vector $\vec{\mathbf{a}}$ in a system with basis vectors $\vec{\mathbf{e}}_1$, $\vec{\mathbf{e}}_2$, $\vec{\mathbf{e}}_3$, defined by the formula

$$\vec{\mathbf{a}} = a^1\,\vec{\mathbf{e}}_1 + a^2\,\vec{\mathbf{e}}_2 + a^3\,\vec{\mathbf{e}}_3 = a^{\alpha}\,\vec{\mathbf{e}}_{\alpha} \tag{3.277}$$

are also called *contravariant coordinates* of this vector. The *covariant coordinates* are the coefficients in the decomposition with the basis vectors $\vec{\mathbf{e}}^1$, $\vec{\mathbf{e}}^2$, $\vec{\mathbf{e}}^3$, i.e., with the reciprocal basis vectors of $\vec{\mathbf{e}}_1$, $\vec{\mathbf{e}}_2$, $\vec{\mathbf{e}}_3$. With the covariant coordinates a_1, a_2, a_3 of the vector $\vec{\mathbf{a}}$

$$\vec{\mathbf{a}} = a_1\,\vec{\mathbf{e}}^1 + a_2\,\vec{\mathbf{e}}^2 + a_3\,\vec{\mathbf{e}}^3 = a_{\alpha}\,\vec{\mathbf{e}}^{\alpha}. \tag{3.278}$$

In the Cartesian coordinate system the covariant and contravariant coordinates of a vector coincide.

2. Representation of Coordinates with Scalar Product
The covariant coordinates of a vector $\vec{\mathbf{a}}$ are equal to the scalar product of this vector with the corresponding basis vectors of the coordinate system:

$$a_1 = \vec{\mathbf{a}}\,\vec{\mathbf{e}}_1, \quad a_2 = \vec{\mathbf{a}}\,\vec{\mathbf{e}}_2, \quad a_3 = \vec{\mathbf{a}}\,\vec{\mathbf{e}}_3. \tag{3.279}$$

The contravariant coordinates of a vector $\vec{\mathbf{a}}$ are equal to the scalar product of this vector with the corresponding basis vectors:

$$a^1 = \vec{\mathbf{a}}\,\vec{\mathbf{e}}^1, \quad a^2 = \vec{\mathbf{a}}\,\vec{\mathbf{e}}^2, \quad a^3 = \vec{\mathbf{a}}\,\vec{\mathbf{e}}^3. \tag{3.280}$$

In Cartesian coordinates (3.279) and (3.280) are coincident:

$$a_x = \vec{\mathbf{a}}\,\vec{\mathbf{i}}, \quad a_y = \vec{\mathbf{a}}\,\vec{\mathbf{j}}, \quad a_z = \vec{\mathbf{a}}\,\vec{\mathbf{k}}. \tag{3.281}$$

Table 3.17 Vector equations

$\vec{\mathbf{x}}$ unknown vector; $\vec{\mathbf{a}}$, $\vec{\mathbf{b}}$, $\vec{\mathbf{c}}$, $\vec{\mathbf{d}}$ given vectors; x, y, z unknown scalars; α, β, γ given scalars

Equation	Solution
1. $\vec{\mathbf{x}}+\vec{\mathbf{a}}=\vec{\mathbf{b}}$	$\vec{\mathbf{x}}=\vec{\mathbf{b}}-\vec{\mathbf{a}}$
2. $\alpha\vec{\mathbf{x}}=\vec{\mathbf{a}}$	$\vec{\mathbf{x}}=\dfrac{\vec{\mathbf{a}}}{\alpha}$
3. $\vec{\mathbf{x}}\,\vec{\mathbf{a}}=\alpha$	Indeterminate equation; considering all vectors $\vec{\mathbf{x}}$ satisfying the equation, with the same initial point, then the endpoints form a plane perpendicular to the vector $\vec{\mathbf{a}}$. Equation 3. is called *the vector equation* of this *plane*.
4. $\vec{\mathbf{x}}\times\vec{\mathbf{a}}=\vec{\mathbf{b}}\quad(\vec{\mathbf{b}}\perp\vec{\mathbf{a}})$	Indeterminate equation; considering all vectors $\vec{\mathbf{x}}$ satisfying the equation, with the same initial point, then the endpoints form a line parallel to $\vec{\mathbf{a}}$. Equation 4. is called the *vector equation* of this *line*.
5. $\begin{cases}\vec{\mathbf{x}}\,\vec{\mathbf{a}}=\alpha\\ \vec{\mathbf{x}}\times\vec{\mathbf{a}}=\vec{\mathbf{b}}\quad(\vec{\mathbf{b}}\perp\vec{\mathbf{a}})\end{cases}$	$\vec{\mathbf{x}}=\dfrac{\alpha\,\vec{\mathbf{a}}+\vec{\mathbf{a}}\times\vec{\mathbf{b}}}{a^2}\qquad(a=\lvert\vec{\mathbf{a}}\rvert)$
6. $\begin{cases}\vec{\mathbf{x}}\,\vec{\mathbf{a}}=\alpha\\ \vec{\mathbf{x}}\,\vec{\mathbf{b}}=\beta\\ \vec{\mathbf{x}}\,\vec{\mathbf{c}}=\gamma\end{cases}$	$\vec{\mathbf{x}}=\dfrac{\alpha(\vec{\mathbf{b}}\times\vec{\mathbf{c}})+\beta(\vec{\mathbf{c}}\times\vec{\mathbf{a}})+\gamma(\vec{\mathbf{a}}\times\vec{\mathbf{b}})}{\vec{\mathbf{a}}\vec{\mathbf{b}}\,\vec{\mathbf{c}}}=\alpha\tilde{\vec{\mathbf{a}}}+\beta\tilde{\vec{\mathbf{b}}}+\gamma\tilde{\vec{\mathbf{c}}}$, where $\tilde{\vec{\mathbf{a}}}$, $\tilde{\vec{\mathbf{b}}}$, $\tilde{\vec{\mathbf{c}}}$ are the reciprocal vectors of $\vec{\mathbf{a}}$, $\vec{\mathbf{b}}$, $\vec{\mathbf{c}}$ (see 3.5.1.6, **1.**, p. 186).
7. $\vec{\mathbf{d}}=x\,\vec{\mathbf{a}}+y\,\vec{\mathbf{b}}+z\,\vec{\mathbf{c}}$	$x=\dfrac{\vec{\mathbf{d}}\,\vec{\mathbf{b}}\,\vec{\mathbf{c}}}{\vec{\mathbf{a}}\vec{\mathbf{b}}\,\vec{\mathbf{c}}},\quad y=\dfrac{\vec{\mathbf{a}}\,\vec{\mathbf{d}}\,\vec{\mathbf{c}}}{\vec{\mathbf{a}}\,\vec{\mathbf{b}}\,\vec{\mathbf{c}}},\quad z=\dfrac{\vec{\mathbf{a}}\,\vec{\mathbf{b}}\,\vec{\mathbf{d}}}{\vec{\mathbf{a}}\,\vec{\mathbf{b}}\,\vec{\mathbf{c}}}$
8. $\vec{\mathbf{d}}=x(\vec{\mathbf{b}}\times\vec{\mathbf{c}})+y\,(\vec{\mathbf{c}}\times\vec{\mathbf{a}})+z\,(\vec{\mathbf{a}}\times\vec{\mathbf{b}})$	$x=\dfrac{\vec{\mathbf{d}}\,\vec{\mathbf{a}}}{\vec{\mathbf{a}}\,\vec{\mathbf{b}}\,\vec{\mathbf{c}}},\quad y=\dfrac{\vec{\mathbf{d}}\,\vec{\mathbf{b}}}{\vec{\mathbf{a}}\,\vec{\mathbf{b}}\,\vec{\mathbf{c}}},\quad z=\dfrac{\vec{\mathbf{d}}\,\vec{\mathbf{c}}}{\vec{\mathbf{a}}\,\vec{\mathbf{b}}\,\vec{\mathbf{c}}}$

3. Representation of the Scalar Product in Coordinates

The determination of the scalar product of two vectors by their contravariant coordinates yields the formula (3.273). The corresponding formula for covariant coordinates is:

$$\vec{\mathbf{a}}\,\vec{\mathbf{b}}=g^{\alpha\beta}\,a_\alpha\,b_\beta, \tag{3.282}$$

where $g^{mn}=\vec{\mathbf{e}}^{\,m}\,\vec{\mathbf{e}}^{\,n}$ are the metric coefficients in the system with the reciprocal vectors. Their relation with the coefficients g_{mn} is

$$g^{mn}=\frac{(-1)^{m+n}\,A^{mn}}{\begin{vmatrix}g_{11}&g_{12}&g_{13}\\ g_{21}&g_{22}&g_{23}\\ g_{31}&g_{32}&g_{33}\end{vmatrix}}, \tag{3.283}$$

where A^{mn} is the subdeterminant of the determinant in the denominator obtained by deleting the row and column of the element g_{mn}.

If the vector $\vec{\mathbf{a}}$ is given by covariant coordinates, and the vector $\vec{\mathbf{b}}$ by contravariant coordinates, then their scalar product is

$$\vec{\mathbf{a}}\,\vec{\mathbf{b}}=a^1\,b_1+a^2\,b_2+a^3\,b_3=a^\alpha\,b_\alpha \tag{3.284}$$

and analogously holds

$$\vec{\mathbf{a}}\,\vec{\mathbf{b}}=a_\alpha\,b^\alpha. \tag{3.285}$$

3.5.1.9 Geometric Applications of Vector Algebra

Table 3.18 demonstrates some geometric applications of vector algebra. Other applications from analytic geometry, such as vector equations of the plane and of the line, are demonstrated in 3.5.1.7, p. 189 and 3.5.3.10, p. 218ff. and on the subsequent pages.

Table 3.18 Geometric application of vector algebra

Determination	**Vector formula**	**Formula with coordinates (in Cartesian coordinates)**
Length of the vector $\vec{\mathbf{a}}$	$a = \sqrt{\vec{\mathbf{a}}^2}$	$a = \sqrt{a_x^2 + a_y^2 + a_z^2}$
Area of the parallelogram determined by the vectors $\vec{\mathbf{a}}$ and $\vec{\mathbf{b}}$	$S = \left\lvert \vec{\mathbf{a}} \times \vec{\mathbf{b}} \right\rvert$	$S = \sqrt{\begin{vmatrix} a_y & a_z \\ b_y & b_z \end{vmatrix}^2 + \begin{vmatrix} a_z & a_x \\ b_z & b_x \end{vmatrix}^2 + \begin{vmatrix} a_x & a_y \\ b_x & b_y \end{vmatrix}^2}$
Volume of the parallelepiped determined by the vectors $\vec{\mathbf{a}}$, $\vec{\mathbf{b}}$, $\vec{\mathbf{c}}$	$V = \left\lvert \vec{\mathbf{a}}\vec{\mathbf{b}}\,\vec{\mathbf{c}} \right\rvert$	$V = \begin{vmatrix} a_x & a_y & a_z \\ b_x & b_y & b_z \\ c_x & c_y & c_z \end{vmatrix}$
Angle between the vectors $\vec{\mathbf{a}}$ and $\vec{\mathbf{b}}$	$\cos\varphi = \dfrac{\vec{\mathbf{a}}\vec{\mathbf{b}}}{\sqrt{\vec{\mathbf{a}}^2\vec{\mathbf{b}}^2}}$	$\cos\varphi = \dfrac{a_x b_x + a_y b_y + a_z b_z}{\sqrt{a_x^2 + a_y^2 + a_z^2}\,\sqrt{b_x^2 + b_y^2 + b_z^2}}$

3.5.2 Analytical Geometry of the Plane

3.5.2.1 Basic Concepts, Coordinate Systems in the Plane

The position of every point P of a plane can be given by an arbitrary *coordinate system*. The numbers determining the position of the point are called *coordinates*. Mostly Cartesian coordinates and polar coordinates are in use.

1. Cartesian or Descartes Coordinates

The Cartesian coordinates of a point P are the signed distances of this point, given in a certain measure, from two *coordinate axes* perpendicular to each other **(Fig. 3.120)**. The intersection point 0 of the coordinate axes is called the *origin*. The horizontal coordinate axis, usually the x*-axis*, is usually called the *axis of abscissae*, the vertical coordinate axis, usually the y*-axis*, is the *axis of ordinates*.

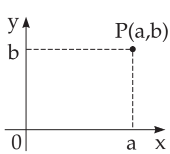


Figure 3.120

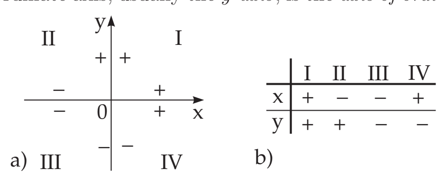


Figure 3.121

The positive direction is given on these axes: on the x-axis usually to the right, on the y-axis upwards. The coordinates of a point P are positive or negative according to which half-axis the projections of the point fall **(Fig. 3.121)**. The coordinates x and y are called the *abscissa* and the *ordinate* of the point P, respectively. The point with abscissa a and ordinate b is denoted by $P(a, b)$. The x, y plane is divided into four *quadrants* I, II, III, and IV by the coordinate axes **(Fig. 3.121,a)**.

2. Polar Coordinates

The polar coordinates of a point P **(Fig. 3.122)** are the *radius* ρ, i.e., the distance of the point from a given point, the *pole* 0, and the *polar angle* φ, i.e., the angle between the line $0P$ and a given oriented half-line passing through the pole, the *polar axis*. The pole is also called the origin. The polar angle is positive if it is measured counterclockwise from the polar axis, otherwise it is negative.

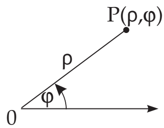


Figure 3.122

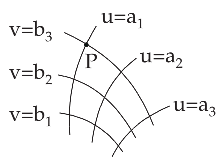


Figure 3.123

3. Curvilinear Coordinate System

This system consists of two one-parameter families of curves in the plane, the family of coordinate curves **(Fig. 3.123)**. Exactly one curve of both families passes through every point of the plane. They intersect each other at this point. The parameters corresponding to this point are its *curvilinear coordinates*. In **Fig. 3.123** the point P has curvilinear coordinates $u = a_1$ and $v = b_3$. In the Cartesian coordinate system the coordinate curves are straight lines parallel to the coordinate axes; in the polar coordinate system the coordinate curves are concentric circles with the center at the pole, and half-lines starting at the pole.

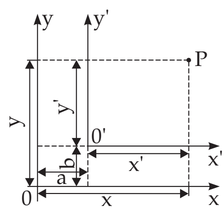


Figure 3.124

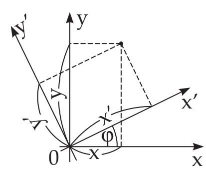


Figure 3.125

3.5.2.2 Coordinate Transformations

Under transformation of a Cartesian coordinate system into another one, the coordinates change according to certain rules.

1. Parallel Translation of Coordinate Axes

The axis of the abscissae is shifted by a, and the axis of the ordinates by b **(Fig. 3.124)**. Suppose a point P has coordinates x, y before the translation, and it has the coordinates x', y' after it. The old coordinates of the new origin $0'$ are a, b. The relations between the old and the new coordinates are the following:

$$x = x' + a, \quad y = y' + b, \tag{3.286a}$$

$$x' = x - a, \quad y' = y - b. \tag{3.286b}$$

2. Rotation of Coordinate Axes

Rotation by an angle φ **(Fig. 3.125)** yields the following changes in the coordinates:

$$\begin{aligned} x' &= x\cos\varphi + y\sin\varphi\,, \\ y' &= -x\sin\varphi + y\cos\varphi\,, \end{aligned} \tag{3.287a}$$

$$\begin{aligned} x &= x'\cos\varphi - y'\sin\varphi\,, \\ y &= x'\sin\varphi + y'\cos\varphi\,. \end{aligned} \tag{3.287b}$$

The coefficient matrix belonging to (3.287a)

$$\mathbf{D} = \begin{pmatrix} \cos\varphi & \sin\varphi \\ -\sin\varphi & \cos\varphi \end{pmatrix} \text{ with } \begin{pmatrix} x' \\ y' \end{pmatrix} = \mathbf{D}\begin{pmatrix} x \\ y \end{pmatrix} \text{ and } \begin{pmatrix} x \\ y \end{pmatrix} = \mathbf{D}^{-1}\begin{pmatrix} x' \\ y' \end{pmatrix}, \tag{3.287c}$$

is called the *rotation matrix*.
In general, the transformation of a cartesian coordinate system into another can be performed in two steps, a translation and a rotation of the coordinate axes.

Remark: With the so-called *coordinate transformation* considered here the coordinate system is transformed, but the represented object rests in its position. In contrast to this with a so-called *geometric transformation* the object is transformed, but the coordinate system remains unchanged in its position. In 3.5.4, p. 229 the rotation of an object is described by

$$\begin{pmatrix} x'_P \\ y'_P \end{pmatrix} = \mathbf{R}\begin{pmatrix} x_P \\ y_P \end{pmatrix}, \tag{3.288}$$

where R is the rotation matrix. Between $\mathbf{D}$ and $\mathbf{R}$ there exists the relation

$$\mathbf{R} = \mathbf{D}^{-1}. \tag{3.289}$$

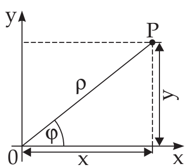


Figure 3.126

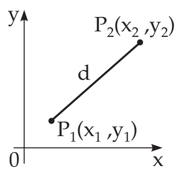


Figure 3.127

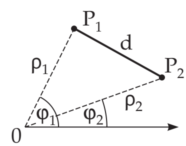


Figure 3.128

3. Transforming Cartesian Coordinates into Polar Coordinates and Conversely

Supposing the origin coincides with the pole and the axis of abscissae coincides with the polar axis **(Fig. 3.126)**, then

$$x = \rho(\varphi)\cos\varphi \quad y = \rho(\varphi)\sin\varphi \quad (-\pi < \varphi \le \pi, \quad \rho \ge 0); \tag{3.290a}$$

$$\rho = \sqrt{x^2 + y^2}, \tag{3.290b}$$

$$\varphi = \begin{cases} \arctan\dfrac{y}{x} + \pi & \text{for } x < 0, \\ \arctan\dfrac{y}{x} & \text{for } x > 0, \\ \dfrac{\pi}{2} & \text{for } x = 0 \text{ and } y > 0, \\ -\dfrac{\pi}{2} & \text{for } x = 0 \text{ and } y < 0, \\ \text{undefined} & \text{for } x = y = 0. \end{cases} \tag{3.290c}$$

3.5.2.3 Special Notations and Points in the Plane

1. Distance Between Two Points

If the two points given in Cartesian coordinates as $P_1\,(x_1, y_1)$ and $P_2\,(x_2, y_2)$ **(Fig. 3.127)**, then their distance is

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}. \tag{3.291}$$

If they are given in polar coordinates as $P_1\,(\rho_1, \varphi_1)$ and $P_2\,(\rho_2, \varphi_2)$ **(Fig. 3.128)**, their distance is

$$d = \sqrt{\rho_1^2 + \rho_2^2 - 2\rho_1\rho_2\cos(\varphi_2 - \varphi_1)}. \tag{3.292}$$

2. Coordinates of Center of Mass

The coordinates (x, y) of the center of mass of a system of material points $M_i\,(x_i, y_i)$ with masses $m_i \ (i = 1, 2, \ldots, n)$ are calculated by the following formula:

$$x = \frac{\sum m_i x_i}{\sum m_i}, \quad y = \frac{\sum m_i y_i}{\sum m_i}. \tag{3.293}$$

3. Division of a Line Segment

1. Division in a Given Ratio The coordinates of the point P with division ratio $\dfrac{\overline{P_1P}}{\overline{PP_2}} = \dfrac{m}{n} = \lambda$ **(Fig. 3.129a)** of the line segment $\overline{P_1P_2}$ are calculated by the formulas

$$x = \frac{nx_1 + mx_2}{n+m} = \frac{x_1 + \lambda x_2}{1+\lambda}, \quad (3.294a) \qquad y = \frac{ny_1 + my_2}{n+m} = \frac{y_1 + \lambda y_2}{1+\lambda}. \quad (3.294b)$$

For the *midpoint* M of the segment $\overline{P_1P_2}$, because of $\lambda = 1$

$$x = \frac{x_1 + x_2}{2}, \quad (3.294c) \qquad y = \frac{y_1 + y_2}{2} \quad (3.294d)$$

holds. The sign of the segments $\overline{P_1P}$ and $\overline{PP_2}$ can be defined. Their signs are positive or negative depending on whether their directions are coincident with $\overline{P_1P_2}$ or not. Then formulas (3.294a,b,c,d) result in a point outside of the segment $\overline{P_1P_2}$ in the case $\lambda < 0$. This is called an *external division*.

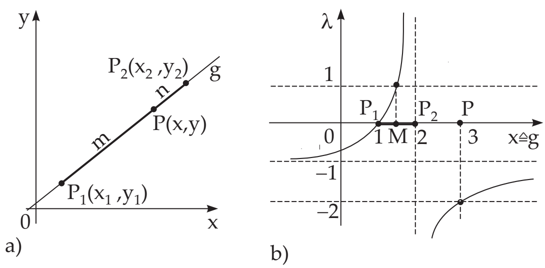


Figure 3.129

If P is inside the segment $\overline{P_1P_2}$, it is called an *internal division*. One defines:
a) $\lambda = 0$ if $P = P_1$,
b) $\lambda = \infty$ if $P = P_2$ and
c) $\lambda = -1$ if P is an infinite or improper point of the line g, i.e., if P is infinitely far from $\overline{P_1P_2}$ on g.
The shape of λ is shown in **Fig. 3.129b**.

■ For a point P, for which P_2 is the midpoint of the segment $\overline{P_1P}$, $\lambda = \dfrac{\overline{P_1P}}{\overline{PP_2}} = -2$ holds.

2. Harmonic Division If the internal and external division of a line segment have the same absolute value $|\lambda|$, it is called *harmonic division*. Denote by P_i and P_a the points of the internal and external division respectively, and by λ_i and λ_a the internal and external rates. Then

$$\frac{\overline{P_1P_i}}{\overline{P_iP_2}} = \lambda_i = \frac{\overline{P_1P_a}}{\overline{P_aP_2}} = -\lambda_a \quad (3.295a) \qquad \text{or} \qquad \lambda_i + \lambda_a = 0. \quad (3.295b)$$

If M denotes the midpoint of the segment $\overline{P_1P_2}$ at a distance b from P_1 **(Fig. 3.130)**, and the distances of P_i and P_a from M are denoted by x_i and x_a, then

$$\frac{b + x_i}{b - x_i} = \frac{x_a + b}{x_i - b} \quad \text{or} \quad \frac{x_i}{b} = \frac{b}{x_a}, \quad \text{i.e.,} \quad x_i x_a = b^2. \tag{3.296}$$

The name *harmonic division* is in connection with the harmonic mean (see 1.2.5.3, p. 20). In **Fig. 3.131** the harmonic division is represented for $\lambda = 5 : 1$, analogously to **Fig. 3.14**. The harmonic mean r of

the segments $\overline{P_1P_i} = p$ and $\overline{P_1P_a} = q$ according to (3.295a) equals in accordance with (1.67b), p. 20, to

$$r = \frac{2pq}{p+q} = 2b \quad (\text{see } \textbf{Fig. 3.132}). \tag{3.297}$$

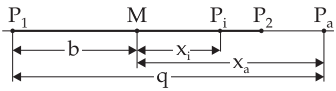


Figure 3.130

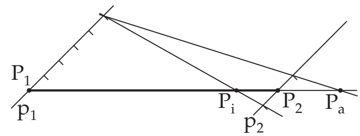


Figure 3.131

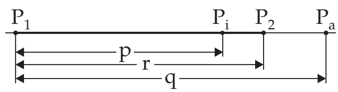
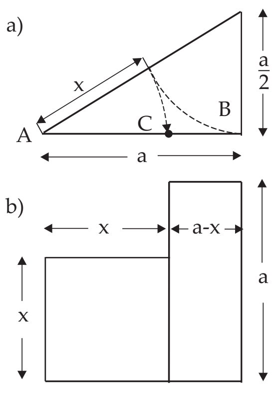


Figure 3.132

3. Golden Section of a segment a is its division into two parts x and $a-x$ such that the part x and the whole segment a have the same ratio as the parts $a-x$ and x:



$$\frac{x}{a} = \frac{a-x}{x}. \tag{3.298a}$$

In this case x is the geometric mean of a and $a-x$ (see also golden section in 1.1.1.2, p. 2), and it holds:

$$x = \sqrt{a(a-x)}, \quad (3.298b) \qquad x = \frac{a(\sqrt{5}-1)}{2} \approx 0.618 \cdot a. \quad (3.298c)$$

The part x of the segment can be geometrically constructed as shown in **Fig. 3.133a**.

Remark 1: The line segment x is also the length of the side of a regular decagon with circum radius a (see also 3.1.5.3, p. 139).

Remark 2: The following geometrical problem produces also the equation of the golden section: Given a rectangle with the constant side length a and a variable side $a-x$. To find is the value x so, that the area $a(a-x)$ of the rectangle is equal to the area x^2 of the square (see **Fig.3.133b**).

Figure 3.133

3.5.2.4 Areas

1. Area of a Convex Polygon

If the vertices are given by $P_1(x_1, y_1)$, $P_2(x_2, y_2)$, ..., $P_n(x_n, y_n)$, then the area is

$$S = \frac{1}{2}\left[(x_1 - x_2)(y_1 + y_2) + (x_2 - x_3)(y_2 + y_3) + \cdots + (x_n - x_1)(y_n + y_1)\right]. \tag{3.299}$$

The formulas (3.299) and (3.300) result in a positive area if the vertices are enumerated counterclockwise, otherwise the area is negative.

2. Area of a Triangle

If the vertices are given by $P_1(x_1, y_1)$, $P_2(x_2, y_2)$, and $P_3(x_3, y_3)$ **(Fig. 3.134)**, then the area can be calculated by the formula

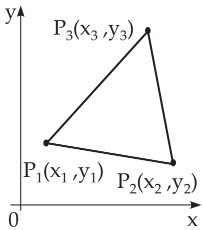


Figure 3.134

$$S = \frac{1}{2}\begin{vmatrix} x_1 & y_1 & 1 \\ x_2 & y_2 & 1 \\ x_3 & y_3 & 1 \end{vmatrix} = \frac{1}{2}\left[x_1(y_2 - y_3) + x_2(y_3 - y_1) + x_3(y_1 - y_2)\right]$$

$$= \frac{1}{2}\left[(x_1 - x_2)(y_1 + y_2) + (x_2 - x_3)(y_2 + y_3) + (x_3 - x_1)(y_3 + y_1)\right]. \quad (3.300)$$

Three points P_1, P_2, P_3 are on the same line if $\begin{vmatrix} x_1 & y_1 & 1 \\ x_2 & y_2 & 1 \\ x_3 & y_3 & 1 \end{vmatrix} = 0$ holds. (3.301)

3.5.2.5 Equation of a Curve

An equation $F(x, y) = 0$ for the coordinates x and y often corresponds to a curve, which has the property that every of its points P satisfies the equation, and conversely, every point whose coordinates satisfy the equation is on the curve. The set of these points is also called the *geometric locus* or simply *locus*. If there is no real point in the plane satisfying the equation $F(x, y) = 0$, then there is no real curve, and one talks about an *imaginary curve*:

■ **A:** $x^2 + y^2 + 1 = 0$,

■ **B:** $y - \ln\left(1 - x^2 - \cosh x\right) = 0$.

The curve corresponding to the equality $F(x, y) = 0$ is called an *algebraic curve* if $F(x, y)$ is a polynomial, and the degree of the polynomial is the *order* or *degree of the curve* (see 2.3.4, p. 65). If the equation of the curve cannot be transformed into the form $F(x, y) = 0$ with a polynomial expression $F(x, y)$, then the curve is called a *transcendental curve*.

The equation of a curve can be defined in the same way in any coordinate system. But from now on, in this book only the Cartesian coordinate system is used, except when stated otherwise.

3.5.2.6 Line

1. Equation of the Line

Every equation that is linear in the coordinates is the equation of a line, and conversely, the equation of every line is a linear equation of the coordinates.

1. General Equation of Line

$$Ax + By + C = 0 \quad (A, B, C \text{ const}). \quad (3.302)$$

For $A = 0$ **(Fig. 3.135)** the line is parallel to the x-axis, for $B = 0$ it is parallel to the y-axis, for $C = 0$ it passes through the origin.

2. Equation of the Line with Slope (or Angular Coefficient) Every line that is not parallel to the y-axis can be represented by an equation written in the form

$$y = kx + b \quad (k, b \text{ const}). \quad (3.303)$$

The quantity k is called the *angular coefficient* or *slope* of the line; it is equal to the tangent of the angle between the line and the positive direction of the x-axis **(Fig. 3.136)**. The line cuts out the segment b from the y-axis. Both the tangent and the value of b can be negative, depending on the position of the line.

3. Equation of a Line Passing Through a Given Point The equation of a line which goes through a given point $P_1(x_1, y_1)$ in a given direction **(Fig. 3.137)** is

$$y - y_1 = k(x - x_1), \quad \text{with} \quad k = \tan\delta. \quad (3.304)$$

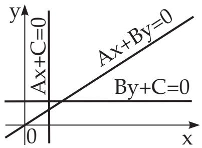


Figure 3.135

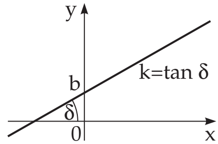


Figure 3.136

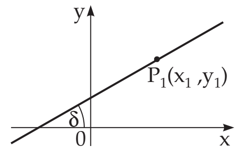


Figure 3.137

4. Equation of a Line Passing Through Two Given Points If two points of the line $P_1\,(x_1, y_1)$, $P_2\,(x_2, y_2)$ are given **(Fig. 3.138)**, then the equation of the line is

$$\frac{y - y_1}{y_2 - y_1} = \frac{x - x_1}{x_2 - x_1}. \tag{3.305}$$

5. Intercept Equation of a Line If a line cuts out the segments a and b from the coordinate axes, considering them with sign, the equation of the line is **(Fig. 3.139)**

$$\frac{x}{a} + \frac{y}{b} = 1. \tag{3.306}$$

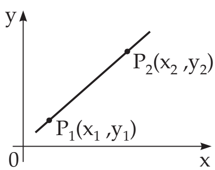


Figure 3.138

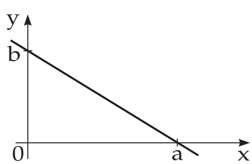


Figure 3.139

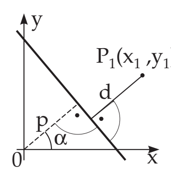


Figure 3.140

6. Normal Form of the Equation of the Line (Hessian Normal Form) With p as the distance of the line from the origin, and with α as the angle between the x-axis and the normal of the line passing through the origin **(Fig. 3.140)**, with $p > 0$, and $0 \le \alpha < 2\pi$, the Hessian normal form is

$$x \cos\alpha + y \sin\alpha - p = 0. \tag{3.307}$$

The *Hessian normal form* can be got from the general equation if multiply (3.302) by the *normalizing factor*

$$\mu = \pm\frac{1}{\sqrt{A^2 + B^2}}. \tag{3.308}$$

The sign of μ must be the opposite to that of C in (3.302).

7. Equation of a Line in Polar Coordinates (Fig. 3.141) With p as the distance of the line from the pole (normal segment from the pole to the line), and with α as the angle between the polar axis and the normal to the line passing through the pole, the equation of the line is

$$\rho = \frac{p}{\cos(\varphi - \alpha)}. \tag{3.309}$$

2. Distance of a Point from a Line

The distance d of a point $P_1\,(x_1, y_1)$ from a line **(Fig. 3.140)** can be got by substituting the coordinates of the point into the left-hand side of the Hessian normal form (3.307):

$$d = x_1 \cos\alpha + y_1 \sin\alpha - p. \tag{3.310}$$

If P_1 and the origin are on different sides of the line, yields $d > 0$, otherwise $d < 0$.

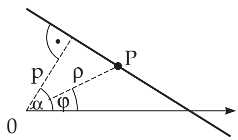


Figure 3.141

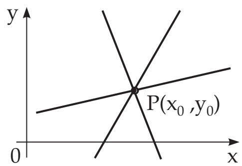


Figure 3.142

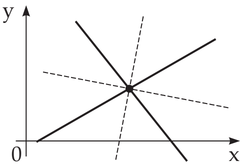


Figure 3.143

3. Intersection Point of Lines

1. Intersection Point of Two Lines In order to get the coordinates (x_0, y_0) of the intersection point of two lines the system of equations given by the equations is to be solved. If the lines are given by the equations

$$A_1x + B_1y + C_1 = 0, \quad A_2x + B_2y + C_2 = 0 \tag{3.311a}$$

then the solution is

$$x_0 = \frac{\begin{vmatrix} B_1 & C_1 \\ B_2 & C_2 \end{vmatrix}}{\begin{vmatrix} A_1 & B_1 \\ A_2 & B_2 \end{vmatrix}}, \quad y_0 = \frac{\begin{vmatrix} C_1 & A_1 \\ C_2 & A_2 \end{vmatrix}}{\begin{vmatrix} A_1 & B_1 \\ A_2 & B_2 \end{vmatrix}}. \tag{3.311b}$$

If $\begin{vmatrix} A_1 & B_1 \\ A_2 & B_2 \end{vmatrix} = 0$ holds, the lines are parallel. If $\frac{A_1}{A_2} = \frac{B_1}{B_2} = \frac{C_1}{C_2}$ holds, the lines are coincident.

2. Pencil of Lines If a third line with equation

$$A_3x + B_3y + C_3 = 0 \tag{3.312a}$$

passes through the intersection point of the first two lines **(Fig. 3.142)**, then the relation

$$\begin{vmatrix} A_1 & B_1 & C_1 \\ A_2 & B_2 & C_2 \\ A_3 & B_3 & C_3 \end{vmatrix} = 0 \tag{3.312b}$$

must be satisfied.
The equation

$$(A_1x + B_1y + C_1) + \lambda\,(A_2x + B_2y + C_2) = 0 \qquad (-\infty < \lambda < +\infty) \tag{3.312c}$$

describes all the lines passing through the intersection point $P_0(x_0, y_0)$ of the two lines (3.311a). By (3.312c) a *pencil of lines* is defined with center $P_0(x_0, y_0)$. If the equations of the first two lines are given in normal form, then $\lambda = \pm 1$ yields the equations of the bisectrices of the angles at the intersection point **(Fig. 3.143)**.

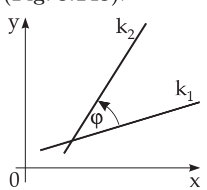


Figure 3.144

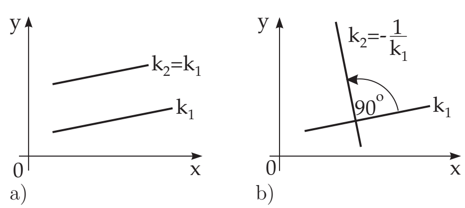


Figure 3.145

4. Angle Between Two Lines

In **Fig. 3.144** there are two intersecting lines. If their equations are given in the general form

$$A_1x + B_1y + C_1 = 0 \quad \text{and} \quad A_2x + B_2y + C_2 = 0, \tag{3.313a}$$

then for the angle φ holds

$$\tan\varphi = \frac{A_1B_2 - A_2B_1}{A_1A_2 + B_1B_2}, \tag{3.313b}$$

$$\cos\varphi = \frac{A_1A_2 + B_1B_2}{\sqrt{A_1^2 + B_1^2}\sqrt{A_2^2 + B_2^2}}, \quad (3.313c) \qquad \sin\varphi = \frac{A_1B_2 - A_2B_1}{\sqrt{A_1^2 + B_1^2}\sqrt{A_2^2 + B_2^2}}. \quad (3.313d)$$

With the slopes k_1 and k_2 of the intersecting lines holds

$$\tan\varphi = \frac{k_2 - k_1}{1 + k_1k_2}, \tag{3.313e}$$

$$\cos\varphi = \frac{1 + k_1k_2}{\sqrt{1 + k_1^2}\sqrt{1 + k_2^2}}, \quad (3.313f) \qquad \sin\varphi = \frac{k_2 - k_1}{\sqrt{1 + k_1^2}\sqrt{1 + k_2^2}}. \quad (3.313g)$$

Here the angle φ is to be considered into the counterclockwise direction from the first line to the second one.

For *parallel lines* **(Fig. 3.145a)** the equalities $\frac{A_1}{A_2} = \frac{B_1}{B_2}$ or $k_1 = k_2$ are valid.

For *perpendicular (orthogonal) lines* **(Fig. 3.145b)** holds $A_1A_2 + B_1B_2 = 0$ or $k_2 = -1/k_1$.

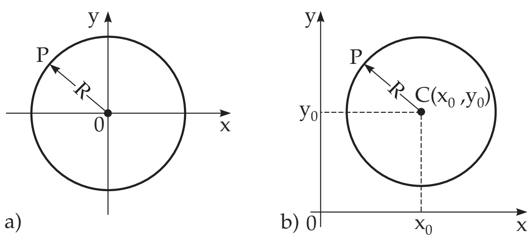


Figure 3.146

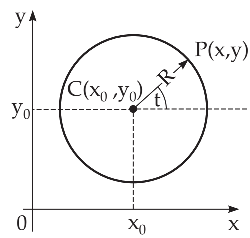


Figure 3.147

3.5.2.7 Circle

1. Definition of the Circle The locus of points at the same given distance from a given point is called a *circle*. The given distance is called the *radius* and the given point is called the *center* of the circle.

2. Equation of the Circle in Cartesian Coordinates The equation of the circle in Cartesian coordinates when its center is at the origin **(Fig. 3.146a)** is

$$x^2 + y^2 = R^2. \tag{3.314a}$$

If the center is at the point $C(x_0, y_0)$ **(Fig. 3.146b)**, then the equation is

$$(x - x_0)^2 + (y - y_0)^2 = R^2. \tag{3.314b}$$

The general equation of second degree

$$ax^2 + 2bxy + cy^2 + 2dx + 2ey + f = 0 \tag{3.315a}$$

is the equation of a circle only if $b = 0$ and $a = c$. In this case the equation can always be transformed into the form

$$x^2 + y^2 + 2mx + 2ny + q = 0. \tag{3.315b}$$

For the radius and the coordinates of the center of the circle the following equalities hold

$$R = \sqrt{m^2 + n^2 - q}\,, \qquad (3.316a) \qquad\qquad x_0 = -m, \quad y_0 = -n. \qquad (3.316b)$$

If $q > m^2 + n^2$ holds, the equation defines an imaginary curve, if $q = m^2 + n^2$ the curve has one single point $P(x_0, y_0)$.

3. Parametric Representation of the Circle

$$x = x_0 + R\cos t, \quad y = y_0 + R\sin t, \qquad (3.317)$$

where t is the angle between the moving radius and the positive direction of the x-axis **(Fig. 3.147)**.

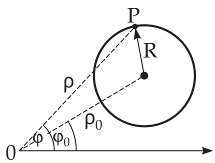


Figure 3.148

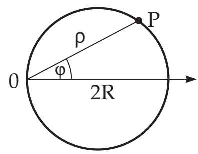


Figure 3.149

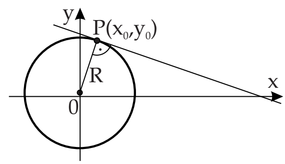


Figure 3.150

4. Equation of the Circle in Polar Coordinates in the general case corresponding to **Fig. 3.148**:

$$\rho^2 - 2\rho\rho_0\cos(\varphi - \varphi_0) + \rho_0^2 = R^2. \qquad (3.318a)$$

If the center is on the polar axis and the circle goes through the origin **(Fig. 3.149)** the equation has the form

$$\rho = 2R\cos\varphi. \qquad (3.318b)$$

5. Tangent of a Circle The equation of the tangent of a circle, given by (3.314a) at the point $P(x_0, y_0)$ **(Fig. 3.150)** has the form

$$xx_0 + yy_0 = R^2. \qquad (3.319)$$

3.5.2.8 Ellipse

1. Elements of the Ellipse In **Fig. 3.151**, $\overline{AB} = 2a$ is the *major axis*, $\overline{CD} = 2b$ is the *minor axis*, A, B, C, D are the *vertices*, F_1, F_2 are the *foci* at a distance $c = \sqrt{a^2 - b^2}$ on both sides from the midpoint, $e = c/a < 1$ is the *numerical eccentricity*, and $p = b^2/a$ is the *semifocal chord*, i.e., the half-length of the chord which is parallel to the minor axis and goes through a focus.

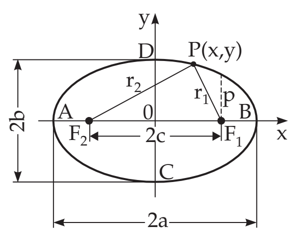


Figure 3.151

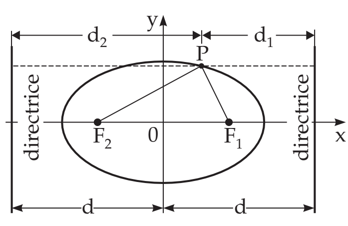


Figure 3.152

2. Equation of the Ellipse If the coordinate axes and the axes of the ellipse are coincident, the equation of the ellipse has the normal form. This equation and the equations in parametric form are

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1, \qquad (3.320a) \qquad\qquad x = a\cos t, \quad y = b\sin t. \qquad (3.320b)$$

For the equation of the ellipse in polar coordinates see 3.5.2.11, **6.**, p. 208.

3. Definition of the Ellipse, Focal Properties The ellipse is the locus of points for which the sum of the distances from two given points, the foci, is a constant, and equal to $2a$. These distances, which are also called the focal radii of the points of the ellipse, can be expressed as a function of the coordinate x from the equalities

$$r_1 = \overline{F_1P} = a - ex, \quad r_2 = \overline{F_2P} = a + ex, \quad r_1 + r_2 = 2a. \tag{3.321}$$

Also here, and in the following formulas in Cartesian coordinates, it is supposed that the ellipse is given in normal form.

4. Directrices of an Ellipse are lines parallel to the minor axis at distance $d = a/e$ from it **(Fig. 3.152)**. Every point $P(x, y)$ of the ellipse satisfies the equalities

$$\frac{r_1}{d_1} = \frac{r_2}{d_2} = e, \tag{3.322}$$

and this property can also be taken as a definition of the ellipse.

5. Diameter of the Ellipse The chords passing through the midpoint of the ellipse are called *diameters of the ellipse*. The midpoint of the ellipse is also the midpoint of the diameter **(Fig. 3.153)**. The locus of the midpoints of all chords parallel to the same diameter is also a diameter; it is called the *conjugate diameter* of the first one. For k and k' as slopes of two conjugate diameters the equality

$$kk' = \frac{-b^2}{a^2} \tag{3.323}$$

holds. If $2a_1$ and $2b_1$ are the lengths of two conjugate diameters and α and β are the acute angles between the diameters and the major axis, where $k = -\tan\alpha$ and $k' = \tan\beta$ hold, then the *Apollonius theorem* holds in the form

$$a_1b_1 \sin(\alpha + \beta) = ab, \qquad a_1^2 + b_1^2 = a^2 + b^2. \tag{3.324}$$

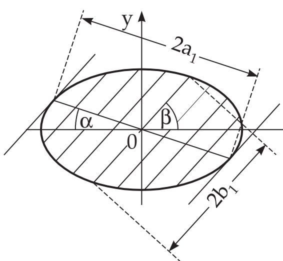


Figure 3.153

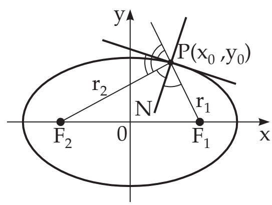


Figure 3.154

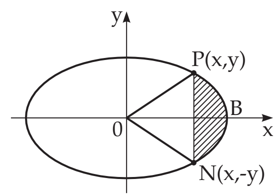


Figure 3.155

6. Tangent of the Ellipse at the point $P(x_0, y_0)$ is given by the equation

$$\frac{xx_0}{a^2} + \frac{yy_0}{b^2} = 1. \tag{3.325}$$

The normal and tangent lines at a point P of the ellipse **(Fig. 3.154)** are bisectors of the interior and exterior angles of the radii connecting the point P with the foci. The line $Ax + By + C = 0$ is a tangent line of the ellipse if the equation

$$A^2a^2 + B^2b^2 - C^2 = 0 \tag{3.326}$$

is satisfied.

7. Radius of Curvature of the Ellipse (Fig. 3.154) If u denotes the angle between the tangent line and the radius vector connecting the point of contact $P(x_0, y_0)$ with a focus, then the radius of

curvature is

$$R = a^2b^2\left(\frac{x_0^2}{a^4}+\frac{y_0^2}{b^4}\right)^{\frac{3}{2}} = \frac{(r_1r_2)^{\frac{3}{2}}}{a\,b} = \frac{p}{\sin^3 u}\,. \tag{3.327}$$

At the vertices A and B **(Fig. 3.151)** and at C and D the radii are $R_A = R_B = \dfrac{b^2}{a} = p$ and $R_C = R_D = \dfrac{a^2}{b}$.

8. Areas of the Ellipse (Fig. 3.155)

a) Ellipse:

$$S = \pi\, a\, b. \tag{3.328a}$$

b) Sector of the Ellipse B0P:

$$S_{\mathrm{B0P}} = \frac{ab}{2}\arccos\frac{x}{a}. \tag{3.328b}$$

c) Segment of the Ellipse PBN:

$$S_{\mathrm{PBN}} = a\,b\arccos\frac{x}{a} - x\,y. \tag{3.328c}$$

9. Arc and Perimeter of the Ellipse The arclength between two points A and B of the ellipse cannot be calculated in an elementary way as for the parabola, but with an incomplete elliptic integral of the second kind $E(k,\varphi)$ (see 8.2.2.2, **2.**, p. 502).
The perimeter of the ellipse (see also 8.2.5, **1.**, 515) can be calculated by a complete elliptic integral of the second kind $E(e) = E\left(e, \dfrac{\pi}{2}\right)$ with the numerical eccentricity $e = \sqrt{a^2-b^2}/a$ and with $\varphi = \dfrac{\pi}{2}$ (for one quadrant of the perimeter), and it is

$$L = 4aE(k,\frac{\pi}{2}) = 4a\int\limits_0^{\pi/2}\sqrt{1-k^2\sin^2\psi}\quad\text{with}\quad k = e = \sqrt{a^2-b^2}/a\,. \tag{3.329a}$$

The calculation of $L = 4aE(E,\pi/2) = 4aE(e)$ can be performed by the help of the following methods:
a) Series expansion

$$L = 4aE(e) = 2\pi a\left[1-\left(\frac{1}{2}\right)^2e^2-\left(\frac{1\cdot 3}{2\cdot 4}\right)^2\frac{e^4}{3}-\left(\frac{1\cdot 3\cdot 5}{2\cdot 4\cdot 6}\right)^2\frac{e^6}{5}-\cdots\right], \tag{3.329b}$$

$$L = \pi(a+b)\left[1+\frac{\lambda^2}{4}+\frac{\lambda^4}{64}+\frac{\lambda^6}{256}+\frac{25\lambda^8}{16384}+\cdots\right]\quad\text{with}\quad\lambda = \frac{(a-b)}{(a+b)}\,. \tag{3.329c}$$

b) Approximate formulas

$$L \approx \pi\left[1,5(a+b)-\sqrt{ab}\right] \tag{3.329d}$$

or

$$L \approx \pi(a+b)\frac{64-3\lambda^4}{64-16\lambda^2}\,. \tag{3.329e}$$

c) Using table 21.9, p. 1103 for the complete elliptic integral of the second kind.

d) Methods of numeric integration to determine the integral in (3.329a).

■ For $a = 1.5$, $b = 1$ one gets the following approximate values for L: according to (3.329e) $L \approx 7.9327$, according to (3.329e) 7.9327, by the help of **table 21.9**, p. 1103 $L \approx 7.94$ (see 8.1.4.3, ■ p. 491) and by numeric integration the more exact value $L \approx 7.932\,721$.

3.5.2.9 Hyperbola

1. Elements of the Hyperbola In **Fig. 3.156** $AB = 2a$ is the real axis; A, B are the *vertices*; 0 the midpoint; F_1 and F_2 are the *foci* at a distance $c > a$ from the midpoint on the *real axis* on both sides; $CD = 2b = 2\sqrt{c^2-a^2}$ is the *imaginary axis*; $p = b^2/a$ the *semifocal chord of the hyperbola*, i.e., the

half-length of the chord which is perpendicular to the real axis and goes through a focus; $e = c/a > 1$ is the *numerical eccentricity*.

2. Equation of the Hyperbola The equation of the hyperbola in *normal form*, i.e., for coincident x and real axes, and the equations in parametric form are

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1, \tag{3.330a}$$

$$x = \pm a \cosh t,\; y = b \sinh t \; (-\infty < t < \infty), \tag{3.330b}$$

$$x = \pm \frac{a}{\cos t},\; y = b \tan t \; (-\frac{\pi}{2} < t < \frac{\pi}{2}). \tag{3.330c}$$

In polar coordinates see 3.5.2.11, **6.**, p. 208.

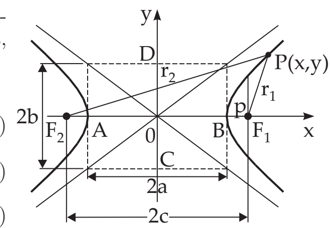


Figure 3.156

3. Definition of the Hyperbola, Focal Properties The hyperbola is the locus of points for which the difference of the distances from two given points, the foci, is a constant $2a$. The points for which $r_1 - r_2 = 2a$ belong to one branch of the hyperbola (in **Fig. 3.156** on the left), the others with $r_2 - r_1 = 2a$ belong to the other branch (in **Fig. 3.156** on the right). These distances, also called the *focal radii*, can be calculated from the formulas

$$r_1 = \pm(ex - a); \quad r_2 = \pm(ex + a); \quad r_2 - r_1 = \pm 2a, \tag{3.331}$$

where the upper sign is valid for the right branch, the lower one for the left branch. Here and in the following formulas for hyperbolas in Cartesian coordinates it is supposed that the hyperbola is given in normal form.

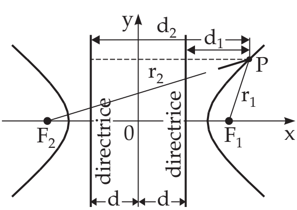


Figure 3.157

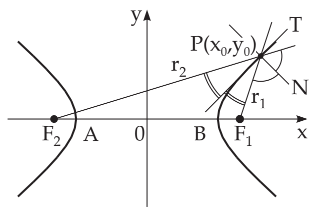


Figure 3.158

4. Directrices of the Hyperbola are the lines perpendicular to the real axis at a distance $d = a/c$ from the midpoint **(Fig. 3.157)**. Every point of the hyperbola $P(x, y)$ satisfies the equalities

$$\frac{r_1}{d_1} = \frac{r_2}{d_2} = e. \tag{3.332}$$

5. Tangent of the Hyperbola at the point $P(x_0, y_0)$ is given by the equation

$$\frac{xx_0}{a_2} - \frac{yy_0}{b_2} = 1. \tag{3.333}$$

The normal and tangent lines of the hyperbola at the point P **(Fig. 3.158)** are bisectors of the interior and exterior angles between the radii connecting the point P with the foci. The line $Ax + By + C = 0$ is a tangent line if the equation

$$A^2a^2 - B^2b^2 - C^2 = 0 \tag{3.334}$$

is satisfied.

6. Asymptotes of the Hyperbola are the lines **(Fig. 3.159)** approached infinitely closely by the branches of the hyperbola for $x \to \infty$. (For the definition of asymptotes see 3.6.1.4, p. 252.) The slopes

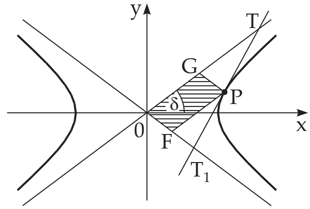


Figure 3.159

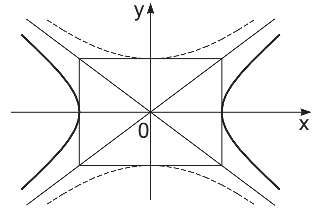


Figure 3.160

of the asymptotes are $k = \pm \tan\delta = \pm b/a$. The equations of the asymptotes are

$$y = \pm\left(\frac{b}{a}\right)x. \tag{3.335}$$

A tangent is intersected by the asymptotes, and they form a *segment of the tangent of the hyperbola*, i.e., the segment $\overline{TT_1}$ **(Fig. 3.159)**. The midpoint of the segment of the tangent is the point of contact P, so $\overline{TP} = \overline{T_1P}$ holds. The area of the triangle $T0T_1$ between the tangent and the asymptotes for any point of contact P is the same, and is

$$S_{TOT_1} = a\,b. \tag{3.336}$$

The area of the parallelogram $0FPG$, determined by the asymptotes and two lines parallel to the asymptotes and passing through the point P, is for any point of contact P

$$S_{OFPG} = \frac{ab}{2}. \tag{3.337}$$

7. Conjugate Hyperbolas (Fig. 3.160) have the equations

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1 \quad \text{and} \quad \frac{y^2}{b^2} - \frac{x^2}{a^2} = 1, \tag{3.338}$$

where the second is represented in **Fig. 3.160** by the dotted line. They have the same asymptotes, hence the real axis of one of them is the imaginary axis of the other one and conversely.

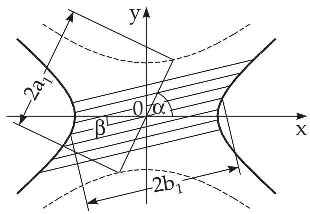


Figure 3.161

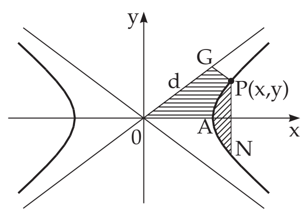


Figure 3.162

8. Diameters of the Hyperbola (Fig. 3.161) are the *chords* between the two branches of the hyperbola passing through the midpoint, which is their midpoint too. Two diameters with slopes k and k' are called conjugate if one of them belongs to a hyperbola and the other one belongs to its conjugate, and $kk' = b^2/a^2$ holds. The midpoints of the chords parallel to a diameter are on its *conjugate diameter* **(Fig. 3.161)**. From two conjugate diameters the one with $|k| < b/a$ intersects the hyperbola. If the lengths of two conjugate diameters are $2a_1$ and $2b_1$, and the acute angles between the diameters and the real axis are α and $\beta < \alpha$, then the equalities

$$a_1^2 - b_1^2 = a^2 - b^2, \quad ab = a_1 b_1 \sin(\alpha - \beta) \tag{3.339}$$

are valid.

9. Radius of Curvature of the Hyperbola At the point $P(x_0, y_0)$ the radius of curvature of the hyperbola is

$$R = a^2b^2\left(\frac{x_0^2}{a^4} + \frac{y_0^2}{b^4}\right)^{3/2} = \frac{{r_1r_2}^{3/2}}{ab} = \frac{p}{\sin^3 u}, \tag{3.340a}$$

where u is the angle between the tangent and the radius vector connecting the point of contact with a focus. At the vertices A and B **(Fig. 3.156)** the radius of curvature is

$$R_A = R_B = p = \frac{b^2}{a}. \tag{3.340b}$$

10. Areas in the Hyperbola (Fig. 3.162)

a) Segment APN:

$$S_{\mathrm{APN}} = xy - ab\ln\left(\frac{x}{a} + \frac{y}{b}\right) = x\,y - a\,b\,\mathrm{Arcosh}\,\frac{x}{a}. \tag{3.341a}$$

b) Area 0APG:

$$S_{0\mathrm{APG}} = \frac{ab}{4} + \frac{ab}{2}\ln\frac{2d}{c}. \tag{3.341b}$$

The line segment $\overline{PG}$ is parallel to the lower asymptote, c is the focal distance and $d = \overline{0G}$.

11. Arc of the Hyperbola The *arclength* between two points A and B of the hyperbola cannot be calculated in an elementary way like the parabola, but it can be calculated by an incomplete elliptic integral of the second kind $E(k, \varphi)$ (see 8.1.4.3,**2.**, p. 502), analogously to the arclength of the ellipse (see 3.5.2.8, ■ p. 201).

12. Equilateral Hyperbola has axes with the same length $a = b$, so its equation is

$$x^2 - y^2 = a^2. \tag{3.342a}$$

The asymptotes of the equilateral hyperbola are the lines $y = \pm x$; they are perpendicular to each other. If the asymptotes coincide with the coordinate axes **(Fig. 3.163)**, then the equation is

$$x\,y = \frac{a^2}{2}. \tag{3.342b}$$

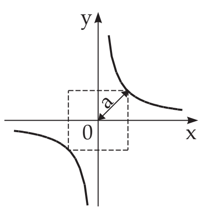


Figure 3.163

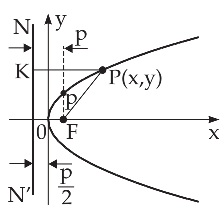


Figure 3.164

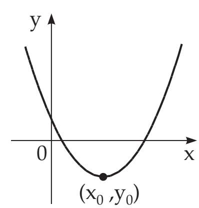


Figure 3.165

3.5.2.10 Parabola

1. Elements of the Parabola In **Fig. 3.164** the x-axis coincides with the *axis of the parabola*, 0 is the *vertex of the parabola*, F is the *focus of the parabola* which is on the x-axis at a distance $p/2$ from the origin, where p is called the *semifocal chord of the parabola.* The *directrix* is denoted by NN'; it is the line perpendicular to the axis of the parabola and intersects the axis at a distance $p/2$ from the origin on the opposite side as the focus. So the semifocal chord is equal to half of the length of the chord which is perpendicular to the axis and passes through the focus. The *numerical eccentricity of the parabola* is equal to 1 (see 3.5.2.11, **4.**, p. 207).

2. Equation of the Parabola If the origin is the vertex of the parabola and the x-axis is the axis of the parabola with the vertex on the left-hand side, then the *normal form of the equation of the parabola* is

$$y^2 = 2p\,x\,. \tag{3.343}$$

For the equation of the parabola in polar coordinates see 3.5.2.11, **6.**, p. 208. For a parabola with vertical axis **(Fig. 3.165)** the equation is

$$y = ax^2 + bx + c. \tag{3.344a}$$

The *parameter* of a parabola given in this form is $p = \dfrac{1}{2|a|}$. (3.344b)

If $a > 0$ holds, the parabola is open up, for $a < 0$ it is open down. The coordinates of the vertex are

$$x_0 = -\frac{b}{2a}\,, \quad y_0 = \frac{4ac - b^2}{4a}\,. \tag{3.344c}$$

3. Properties of the Parabola (Definition of the Parabola) The parabola is the locus of points $P(x, y)$ whose distance from a given point, the focus, is equal to its distance from a given line, the directrix **(Fig. 3.164)**. Here and in the following formulas in Cartesian coordinates the normal form of the equation of the parabola is supposed. Then holds the equation

$$\overline{PF} = \overline{PK} = x + \frac{p}{2}\,, \tag{3.345}$$

where PF is the radius vector whose initial point is at the focus and endpoint is a point of the parabola.

4. Diameter of the Parabola is a line which is parallel to the axis of the parabola **(Fig. 3.166)**. A *diameter of the parabola* halves the chords which are parallel to the tangent line belonging to the endpoint of the diameter **(Fig. 3.166)**. With slope k of the chords the equation of the diameter is

$$y = \frac{p}{k}\,. \tag{3.346}$$

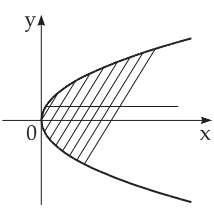


Figure 3.166

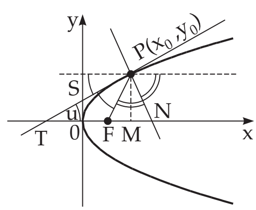


Figure 3.167

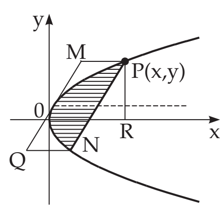


Figure 3.168

5. Tangent of the Parabola (Fig.3.167) The equation of the tangent of the parabola at the point $P(x_0, y_0)$ is

$$yy_0 = p\,(x + x_0)\,. \tag{3.347}$$

Tangent and normal lines are bisectors of the angles between the radius starting at the focus and the diameter starting at the point of contact. The tangent at the vertex, i.e., the y-axis, halves the segment of the tangent line between the point of contact and its intersection point with the axis of the parabola, the x-axis:

$$\overline{TS} = \overline{SP}, \quad \overline{T0} = \overline{0M} = x_0, \quad \overline{TF} = \overline{FP}. \tag{3.348}$$

A line with equation $y = kx + b$ is a tangent line of the parabola if

$$p = 2\,b\,k. \tag{3.349}$$

6. Radius of Curvature of the Parabola at the point $P(x_0, y_0)$ with l_n as the length of the normal $\overline{PN}$ **(Fig. 3.167)** is

$$R = \frac{(p + 2x_1)^{3/2}}{\sqrt{p}} = \frac{p}{\sin^3 u} = \frac{l_n^3}{p^2} \tag{3.350a}$$

and at the vertex 0 it is

$$R = p. \tag{3.350b}$$

7. Areas in the Parabola (Fig. 3.168)

a) Parabolic Segment P0N:

$$S_{0\mathrm{PN}} = \frac{2}{3} S_{\mathrm{MQNP}} \quad (\text{MQNP is a parallelogram}). \tag{3.351a}$$

b) Area 0PR (Area under the Parabola Curve):

$$S_{0PR} = \frac{2xy}{3}. \tag{3.351b}$$

8. Length of Parabolic Arc from the vertex 0 to the point $P(x, y)$

$$l_{0P} = \frac{p}{2}\left[\sqrt{\frac{2x}{p}\left(1 + \frac{2x}{p}\right)} + \ln\left(\sqrt{\frac{2x}{p}} + \sqrt{1 + \frac{2x}{p}}\right)\right] \tag{3.352a}$$

$$= -\sqrt{x\left(x + \frac{p}{2}\right)} + \frac{p}{2}\operatorname{Arsinh}\sqrt{\frac{2x}{p}}. \tag{3.352b}$$

For small values of $\frac{x}{y}$ the following approximation can be used:

$$l_{0P} \approx y\left[1 + \frac{2}{3}\left(\frac{x}{y}\right)^2 - \frac{2}{5}\left(\frac{x}{y}\right)^4\right]. \tag{3.352c}$$

3.5.2.11 Quadratic Curves (Curves of Second Order or Conic Sections)

1. General Equation of Quadratic Curves (Curves of Second Order or Degree)

The ellipse, its special case, the circle, the hyperbola, the parabola or two lines as a singular conic section are defined by the general equation of a quadratic curve (curve of second order)

$$a\,x^2 + 2\,b\,x\,y + c\,y^2 + 2\,d\,x + 2\,e\,y + f = 0. \tag{3.353a}$$

This equation can be reduced to the normal form with the help of the coordinate transformations given in **Tables 3.19** and **3.20**.

Remark 1: The coefficients in (3.353a) are not the parameters of the special conic sections.

Remark 2: If two coefficients (a and b or b and c) are equal to zero, the required coordinate transformation is reduced to a translation of the coordinate axes.
The equation $cy^2 + 2dx + 2ey + f = 0$ can be written in the form $(y - y_0)^2 = 2p(x - x_0)$;
the equation $ax^2 + 2dx + 2ey + f = 0$ can be written in the form $(x - x_0)^2 = 2p\,(y - y_0)$.

2. Invariants of Quadratic Curves

are the three quantities

$$\Delta = \begin{vmatrix} a & b & d \\ b & c & e \\ d & e & f \end{vmatrix}, \quad \delta = \begin{vmatrix} a & b \\ b & c \end{vmatrix}, \quad S = a + c. \tag{3.353b}$$

They do not change during a rotation of the coordinate system, i.e., if after a coordinate transformation

Table 3.19 Equation of curves of second order. Central curves ($\delta \neq 0$)[1]

Quantities δ and Δ			**Shape of the curve**
Central curves $\delta \neq 0$	$\delta > 0$	$\Delta \neq 0$	Ellipse a) for $\Delta \cdot S < 0$: real, b) for $\Delta \cdot S > 0$: imaginary[2]
		$\Delta = 0$	A pair of imaginary[2] lines with real common point
	$\delta < 0$	$\Delta \neq 0$	Hyperbola
		$\Delta = 0$	A pair of intersecting lines

Required coordinate transformations	**Normal form of the equation after the transformation**
1. Translation of the origin to the center of the curve, whose coordinates are $x_0 = \dfrac{be - cd}{\delta}, \quad y_0 = \dfrac{bd - ae}{\delta}.$	$a'x'^2 + c'y'^2 + \dfrac{\Delta}{\delta} = 0$
2. Rotation of the coordinate axes by the angle α with $\tan 2\alpha = \dfrac{2b}{a - c}.$ The sign of $\sin 2\alpha$ must coincide with the sign of $2b$. Here the slope of the new x'-axis is $k = \dfrac{c - a + \sqrt{(c-a)^2 + 4b^2}}{2b}.$	$a' = \dfrac{a + c + \sqrt{(a-c)^2 + 4b^2}}{2}$ $c' = \dfrac{a + c - \sqrt{(a-c)^2 + 4b^2}}{2}$ (a' and c' are the roots of the quadratic equation $u^2 - Su + \delta = 0$).

[1] Δ, δ and S are numbers given in (3.353b).
[2] The equation of the curve corresponds to an imaginary curve.

the equation of the curve has the form

$$a'x'^2 + 2b'x'y' + c'y'^2 + 2d'x' + 2e'y' + f' = 0, \tag{3.353c}$$

then the calculation of these three quantities Δ, δ, and S with the new constants will yield the same values.

3. Shape of the Quadratic Curves (Conic Sections)

If a right double circular cone is intersected by a plane, the result is a conic section. If the plane does not pass through the vertex of the cone, the result is a hyperbola, a parabola, or an ellipse depending on whether the plane is parallel to two, one, or none of the generators of the cone. If the plane goes through the vertex, the result is a *singular conic section* with $\Delta = 0$. As a conic section of a cylinder, i.e., a *singular cone* whose vertex is at infinity, yield parallel lines. The shape of a conic section can be determined with the help of **Tables 3.19** and **3.20**.

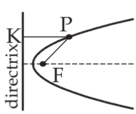


Figure 3.169

4. General Properties of Curves of Second Degree

The locus of every point P **(Fig. 3.169)** with constant ratio e of the distance to a fixed point F, the focus, and the distance from a given line, the directrix, is a curve of second order with *numerical eccentricity* e. For $e < 1$ it is an ellipse, for $e = 1$ it is a parabola, for $e > 1$ it is a hyperbola.

Table 3.20 Equations of curves of second order. Parabolic curves ($\delta = 0$)

Quantity δ and Δ		Shape of the curve
Parabolic curves[1], $\delta = 0$	$\Delta \neq 0$	Parabola
	$\Delta = 0$	Two lines: Parallel lines for $d^2 - af > 0$, Double line for $d^2 - af = 0$, Imaginary[2] lines for $d^2 - af < 0$.

Required coordinate transformation	Normal form of the equation after the transformation
$\Delta \neq 0$ 1. Translation of the origin to the vertex of the parabola whose coordinates x_0 and y_0 are defined by the equations $ax_0 + by_0 + \frac{ad+be}{S} = 0$ and $\left(d + \frac{dc - be}{S}\right) x_0 + \left(e + \frac{ae - bd}{S}\right) y_0 + f = 0\,.$ 2. Rotation of the coordinate axes by the angle α with $\tan\alpha = -\frac{a}{b}$; the sign of $\sin\alpha$ must differ from the sign of a.	$y'^2 = 2px'$ $p = \frac{ae - bd}{S\sqrt{a^2 + b^2}}$
$\Delta = 0$ Rotation of the coordinate axes by the angle α with $\tan\alpha = -\frac{a}{b}$; the sign of $\sin\alpha$ must differ from the sign of a.	$Sy'^2 + 2\frac{ad+be}{\sqrt{a^2+b^2}}y' + f = 0$ can be transformed into the form $(y' - y'_0)(y' - y'_1) = 0$.

[1] In the case of $\delta = 0$ it is supposed that none of the coefficients a, b, c is equal to zero.
[2] The equation of the curve corresponds to an imaginary curve.

5. Determination of a Curve Through Five Points

There is one and only one curve of second degree passing through five given points of a plane. If three of these points are on the same line, then it is a *singular* or *degenerate conic section.*

6. Polar Equation of Curves of Second Degree

All curves of second degree can be described by the polar equation

$$\rho = \frac{p}{1 + e\cos\varphi}, \tag{3.354}$$

where p is the semifocal chord and e is the eccentricity. Here the pole is at the focus, while the polar axis is directed from the focus to the closer vertex. For the hyperbola this equation defines only one branch.

3.5.3 Analytical Geometry of Space

3.5.3.1 Basic Concepts

1. Coordinates and Coordinate Systems

Every point P in space can be determined by a coordinate system. The directions of the coordinate lines are given by the directions of the unit vectors. In **Fig. 3.170a** the relations of a Cartesian coordinate system are represented. There are to distinguish right-angled and oblique coordinate systems where the unit vectors are perpendicular or oblique to each other. Another important difference is whether it is a right-handed or a left-handed coordinate system.

The most common spatial coordinate systems are the Cartesian coordinate system, the spherical polar coordinate system, and the cylindrical polar coordinate system.

2. Right- and Left-Handed Systems

Depending on the successive order of the positive coordinate directions there are to distinguish *right systems* and *left systems* or *right-handed* and *left-handed coordinate systems*. A right system has for instance three non-coplanar unit vectors with indices in alphabetical order $\vec{\mathbf{e}}_i, \vec{\mathbf{e}}_j, \vec{\mathbf{e}}_k$. These vectors form a right-handed system when an observer, looking on the $\vec{\mathbf{e}}_i, \vec{\mathbf{e}}_j$ plane and at the same time into the direction of $\vec{\mathbf{e}}_k$, can rotate $\vec{\mathbf{e}}_i$ into $\vec{\mathbf{e}}_j$ along the smallest angle, i.e., clockwise; looking from the cusp of the vector $\vec{\mathbf{e}}_k$ into the direction of the $\vec{\mathbf{e}}_i, \vec{\mathbf{e}}_j$ plane he can rotate $\vec{\mathbf{e}}_i$ into $\vec{\mathbf{e}}_j$ counterclockwise. A left system consequently requires the opposite rotation in both cases. The alphabetical order of rotation is represented symbolically in **Fig. 3.34**, p. 143 where the notations a, b, c are substituted for the indices i, j, k.

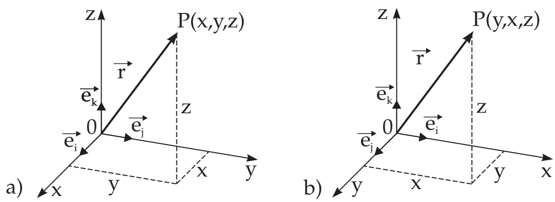


Figure 3.170

Right- and left-handed systems can be transformed into each other by interchanging two unit vectors. The interchange of two unit vectors changes its *orientation*: A right system becomes a left system, and conversely, a left system becomes a right system.

A very important way to interchange vectors is the *cyclic permutation*, where the orientation remains unchanged. As in **Fig. 3.34** the interchange the vectors of a right system by cyclic permutation yields a rotation in a counterclockwise direction, i.e., according to the scheme ($i \to j \to k \to i, j \to k \to i \to j, k \to i \to j \to k$). In a left system the interchange of the vectors by cyclic permutation follows a clockwise rotation, i.e., according to the scheme ($i \to k \to j \to i, k \to j \to i \to k, j \to i \to k \to j$).

A right system is not superposable on a left system.

The reflection of a right system with respect to the origin is a left system (see 4.3.5.1, p. 288).

■ **A:** The Cartesian coordinate system with coordinate axes x, y, z is a right system **(Fig. 3.170a)**.

■ **B:** The Cartesian coordinate system with coordinate axes x, z, y is a left system **(Fig. 3.170b)**.

■ **C:** From the right system $\vec{\mathbf{e}}_i, \vec{\mathbf{e}}_j, \vec{\mathbf{e}}_k$ one gets the left system $\vec{\mathbf{e}}_i, \vec{\mathbf{e}}_k, \vec{\mathbf{e}}_j$ by interchanging the vectors $\vec{\mathbf{e}}_j$ and $\vec{\mathbf{e}}_k$.

■ **D:** By cyclic permutation follows from the right system $\vec{\mathbf{e}}_i, \vec{\mathbf{e}}_j, \vec{\mathbf{e}}_k$ the right system $\vec{\mathbf{e}}_j, \vec{\mathbf{e}}_k, \vec{\mathbf{e}}_i$ and from this one $\vec{\mathbf{e}}_k, \vec{\mathbf{e}}_i, \vec{\mathbf{e}}_j$, a right system again.

Table 3.21 Coordinate signs in the octants

Octant	I	II	III	IV	V	VI	VII	VIII
x	+	−	−	+	+	−	−	+
y	+	+	−	−	+	+	−	−
z	+	+	+	+	−	−	−	−

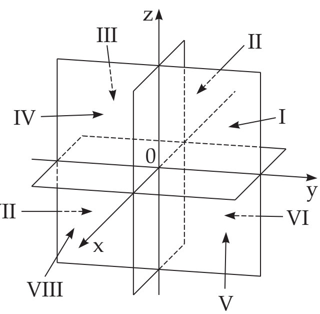


Figure 3.171

3. Cartesian Coordinates

of a point P are its distances from three mutually orthogonal planes in a certain measuring unit, with given signs. They represent the projections of the radius vector $\vec{\mathbf{r}}$ of the point P (see 3.5.1.1, **6.**, p. 182) onto three mutually perpendicular coordinate axes **(Fig. 3.170)**. The intersection point of the planes O, which is the intersection point of the axes too, is called the *origin*. The coordinates x, y, and z are called the *abscissa*, *ordinate*, and *applicate*. The written form $P(a, b, c)$ means that the point P has coordinates $x = a$, $y = b$, $z = c$. The signs of the coordinates are determined by the octant where the point P lies **(Fig. 3.171, Table 3.21)**.

In a right-handed Cartesian coordinate system **(Fig. 3.170a)** for orthogonal unit vectors given in the order $\vec{\mathbf{e}}_i, \vec{\mathbf{e}}_j, \vec{\mathbf{e}}_k$ the equalities

$$\vec{\mathbf{e}}_i \times \vec{\mathbf{e}}_j = \vec{\mathbf{e}}_k\,, \quad \vec{\mathbf{e}}_j \times \vec{\mathbf{e}}_k = \vec{\mathbf{e}}_i\,, \quad \vec{\mathbf{e}}_k \times \vec{\mathbf{e}}_i = \vec{\mathbf{e}}_j\,, \tag{3.355a}$$

hold, i.e., the *right-hand law* is valid (see 3.5.1.5, p. 184). The three formulas transform into each other under *cyclic permutations* of the unit vectors.

In a left-handed Cartesian coordinate system **(Fig. 3.170b)** the equations

$$\vec{\mathbf{e}}_i \times \vec{\mathbf{e}}_j = -\vec{\mathbf{e}}_k\,, \quad \vec{\mathbf{e}}_j \times \vec{\mathbf{e}}_k = -\vec{\mathbf{e}}_i\,, \quad \vec{\mathbf{e}}_k \times \vec{\mathbf{e}}_i = -\vec{\mathbf{e}}_j \tag{3.355b}$$

are valid. The negative sign of the vector product arises from the left-handed order of the unit vectors, see **Fig. 3.170b**, i.e., from their clockwise arrangement.

Notice that in both cases the equations

$$\vec{\mathbf{e}}_i \times \vec{\mathbf{e}}_i = \vec{\mathbf{e}}_j \times \vec{\mathbf{e}}_j = \vec{\mathbf{e}}_k \times \vec{\mathbf{e}}_k = \vec{\mathbf{0}} \tag{3.355c}$$

are valid. Usually one works with right-handed coordinate systems; the formulas do not depend on this choice. In geodesy usually left-handed coordinate systems are in use (see 3.2.2.1, p. 144).

4. Coordinate Surfaces and Coordinate Curves

Coordinate Surfaces have one constant coordinate. In a Cartesian coordinate system they are planes parallel to the plane of other two coordinate axes. By the three coordinate surfaces $x = 0$, $y = 0$, and $z = 0$ three-dimensional space is divided into eight octants **(Fig. 3.171)**. *Coordinate lines* or *coordinate curves* are curves with one changing coordinate while the others are constants. In Cartesian systems they are lines parallel to the coordinate axes. The coordinate surfaces intersect each other in a coordinate line.

3.5.3.2 Spatial Coordinate Systems

1. Curvilinear Three-Dimensional Coordinate Systems

arise if three families of surfaces are given such that for any point of the space there is exactly one surface from every system passing through it. The position of a point will be given by the parameter values of the surfaces passing through it. The most often used curvilinear coordinate systems are the cylindrical polar and the spherical polar coordinate systems.

2. Cylindrical Polar Coordinates (Fig. 3.172)

consist of:
• the polar coordinates ρ and φ of the projection of the point P to the x, y plane and
• the applicate z of the point P.
The coordinate surfaces of a cylindrical polar coordinate system are:
• The cylinder surfaces with radius $\varrho =$ const, and the z-axis as axis of the cylinder,
• the half-planes starting from the z-axis, $\varphi =$ const and
• the planes being perpendicular to the z-axis, $z =$ const.
The intersection curves of these coordinate surfaces are the coordinate curves.

The transformation formulas between the Cartesian coordinate system and the cylindrical polar coordinate system are (see also **Table 3.22**):

$$x = \varrho\cos\varphi, \quad y = \varrho\sin\varphi, \quad z = z; \tag{3.356a}$$

$$\varrho = \sqrt{x^2+y^2}, \quad \varphi = \arctan\frac{y}{x} = \arcsin\frac{y}{\varrho} \quad \text{for} \quad x > 0. \tag{3.356b}$$

For the required distinction of cases with respect to φ see (3.290c), p. 192.

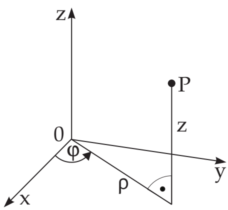


Figure 3.172

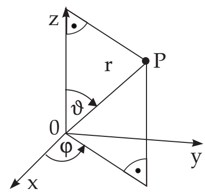


Figure 3.173

3. Spherical Coordinates or Spherical Polar Coordinates (3.173)

contain:
• The length r of the radius vector $\vec{\mathbf{r}}$ of the point P,
• the angle ϑ between the z-axis and the radius vector $\vec{\mathbf{r}}$ and
• the angle φ between the x-axis and the projection of $\vec{\mathbf{r}}$ on the x, y plane.

The positive directions **(Fig. 3.173)** here are for $\vec{\mathbf{r}}$ from the origin to the point P, for ϑ from the z-axis to $\vec{\mathbf{r}}$, and for φ from the x-axis to the projection of $\vec{\mathbf{r}}$ to the x, y plane. With the values $0 \le r < \infty$, $0 \le \vartheta \le \pi$, and $-\pi < \varphi \le \pi$ every point of space can be described.

Coordinate surfaces are:
• Spheres with the origin 0 as center and with radius $r =$ const,
• circular cones with $\vartheta =$ const, with vertex at the origin, and the z-axis as the axis and
• closed half-planes starting at the z-axis with $\varphi =$ const.
The intersection curves of these surfaces are the coordinate curves.

4. Relations between Cartesian, Cylindrical and Spherical Polar Coordinates

■ The transformation formulas between Cartesian coordinates and spherical polar coordinates (see also **Table 3.22**) are:

$$x = r\sin\vartheta\cos\varphi, \quad y = r\sin\vartheta\sin\varphi, \quad z = r\cos\vartheta, \tag{3.357a}$$

$$r = \sqrt{x^2+y^2+z^2}\,, \quad \vartheta = \arctan\frac{\sqrt{x^2+y^2}}{z}\,, \quad \varphi = \arctan\frac{y}{x}\,. \tag{3.357b}$$

For the required distinction of cases with respect to φ see (3.290c), p. 192.

Table 3.22 Relations between Cartesian, cylindrical, and spherical polar coordinates

Cartesian coordinates	Cylindrical polar coordinates	Spherical polar coordinates
$x =$	$= \varrho\cos\varphi$	$= r\sin\vartheta\cos\varphi$
$y =$	$= \varrho\sin\varphi$	$= r\sin\vartheta\sin\varphi$
$z =$	$= z$	$= r\cos\vartheta$
$\sqrt{x^2+y^2}$	$= \varrho$	$= r\sin\vartheta$
$\arctan\dfrac{y}{x}$	$= \varphi$	$= \varphi$
$= z$	$= z$	$= r\cos\vartheta$
$\sqrt{x^2+y^2+z^2}$	$= \sqrt{\varrho^2+z^2}$	$= r$
$\arctan\dfrac{\sqrt{x^2+y^2}}{z}$	$= \arctan\dfrac{\varrho}{z}$	$= \vartheta$
$\arctan\dfrac{y}{x}$	$= \varphi$	$= \varphi$

5. Direction in Space

A direction in the space can be determined by a unit vector $\vec{\mathbf{t}}^0$ (see 3.5.1.1, **6.**, p. 181) whose coordinates are the *direction cosines*, i.e., the cosines of the angles $\alpha_0, \beta_0, \gamma_0$ between the vector and the positive coordinate axes (**Fig. 3.174**)

$$l = \cos\alpha_0, \quad m = \cos\beta_0, \quad n = \cos\gamma_0, \quad l^2 + m^2 + n^2 = 1. \tag{3.358a}$$

The angle φ between two directions given by their direction cosines l_1, m_1, n_1 and l_2, m_2, n_2 can be calculated by the formula

$$\cos\varphi = l_1 l_2 + m_1 m_2 + n_1 n_2. \tag{3.358b}$$

Two directions are perpendicular to each other if

$$l_1 l_2 + m_1 m_2 + n_1 n_2 = 0. \tag{3.358c}$$

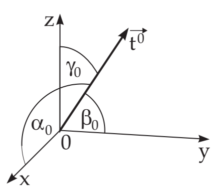


Figure 3.174

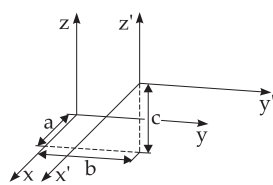


Figure 3.175

3.5.3.3 Transformation of Orthogonal Coordinates

1. Parallel Translation

If the original coordinates are x, y, z, and a, b, c denote the coordinates of the origin of the new coordinate system in the old system (**Fig. 3.175**), then the new coordinates x', y', z' satisfy the relations

$$x = x' + a, \quad y = y' + b, \quad z = z' + c, \quad x' = x - a, \quad y' = y - b, \quad z' = z - c. \tag{3.359}$$

2. Rotation of the Coordinate System

The relation between the original coordinates x, y, z and the new coordinates x', y', z' after rotation is given by

$$\underline{\mathbf{x}}' = \begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = \mathbf{D}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \mathbf{D}\underline{\mathbf{x}}\,. \tag{3.360a}$$

D is called the *rotation matrix of the coordinate system.* The special rotation matrix

$$\mathbf{D}_x(\alpha) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha & \sin\alpha \\ 0 & -\sin\alpha & \cos\alpha \end{pmatrix} \tag{3.360b}$$

describes rotations of the x, y, z system around the x-axis by the angel α. In analogy rotations of the x, y, z system around the y-axis by the angel β or around the z-axis by the angel γ describe the following rotation matrices:

$$\mathbf{D}_y(\beta) = \begin{pmatrix} \cos\beta & 0 & -\sin\beta \\ 0 & 1 & 0 \\ \sin\beta & 0 & \cos\beta \end{pmatrix}, \quad (3.360c) \qquad \mathbf{D}_z(\gamma) = \begin{pmatrix} \cos\gamma & \sin\gamma & 0 \\ -sin\gamma & \cos\gamma & 0 \\ 0 & 0 & 1 \end{pmatrix}. \quad (3.360d)$$

Remark: The rotation of the x, y, z system around an arbitrary axis through the origin can be described with the help of direction cosines (see 3.5.3.4), Cardan angels (3.5.3.5) or Euler angels (3.5.3.6).

3. Rotation of an Object

In geometry two types of transformations are distinguished (see also 3.5.4, p. 229):

a) coordinate transformations (the coordinate system is transformed), and

b) geometric transformations (the position of a geometric object is changed in a fixed coordinate system).

Consequently, at the rotation around an arbitrary axis through the origin the following rotations are distinguished

a) rotation of the coordinate system (see (3.360a-d)) and

b) rotation of an object in a fixed coordinate system. In this case:

$$\underline{\mathbf{x}}'_P = \mathbf{R}\underline{\mathbf{x}}_P \quad (3.361a) \qquad \mathbf{R} = \mathbf{D}^{-1} = \mathbf{D}^{\mathrm{T}}. \quad (3.361b)$$

Here the formulas (3.361a,b) describe the relation between the coordinates x_P, y_P, z_P of the initial position of the object and its coordinates x'_P, y'_P, z'_P after the rotation. **R** is called the *rotation matrix of the object.*

Remarks:

1. Rotations around an arbitrary axis through the origin will have a suitable description with quaternions (see 4.4.2.5, p. 297).

2. Rotations around an arbitrary axis not running through the origin, will be discussed in the example on p. 235.

3.5.3.4 Rotations with Direction Cosines

1. Rotation of the Coordinate Axis

If the direction cosines of the new axes are given as in **Table 3.23**, see also **(Fig. 3.176)**, then for the new and old coordinates

$$\begin{aligned} x' &= l_1x + m_1y + n_1z\,, \\ y' &= l_2x + m_2y + n_2z\,, \\ z' &= l_3x + m_3y + n_3z\,; \end{aligned} \quad (3.362a) \qquad \begin{aligned} x &= l_1x' + l_2y' + l_3z'\,, \\ y &= m_1x' + m_2y' + m_3z'\,, \\ z &= n_1x' + n_2y' + n_3z' \end{aligned} \quad (3.362b)$$

holds. The coefficient matrix of the system (3.362a), which is called the *rotation matrix* **D**, and the *determinant of the transformation* Δ are

$$\mathbf{D} = \begin{pmatrix} l_1 & l_2 & l_3 \\ m_1 & m_2 & m_3 \\ n_1 & n_2 & n_3 \end{pmatrix}, \quad (3.362c) \qquad \det\mathbf{D} = \Delta = \begin{vmatrix} l_1 & l_2 & l_3 \\ m_1 & m_2 & m_3 \\ n_1 & n_2 & n_3 \end{vmatrix}. \quad (3.362d)$$

The following relations are valid:

$$\begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = \mathbf{D} \begin{pmatrix} x \\ y \\ z \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \mathbf{D}^{-1} \begin{pmatrix} x' \\ y' \\ z' \end{pmatrix}. \tag{3.362e}$$

Table 3.23 Notation for the direction cosines under coordinate transformation

With respect to the the old axes	Direction cosine of the new axes		
	x'	y'	z'
x	l_1	l_2	l_3
y	m_1	m_2	m_3
z	n_1	n_2	n_3

2. Properties of the Transformation Determinant

a) $\Delta = \pm 1$, with a positive sign if it remains left- or right-handed, as it was, and with negative sign if it changes its orientation.
b) The sum of the squares of the elements of a row or column is always equal to one.
c) The sum of the products of the corresponding elements of two different rows or columns is equal to zero (see 4.1.4, **9.**, p. 275).
d) Every element can be written as the product of $\Delta = \pm 1$ and its adjoint (see 4.2.1, p. 278).

3. Scalar Invariant

This is a scalar which keeps its value during translation and rotation. The scalar product of two vectors is a *scalar invariant* (see 3.5.1.5, **3.**, p. 185).

■ **A:** The components of a vector $\vec{\mathbf{a}} = \{a_1, a_2, a_3\}$ are not scalar invariants, because they change their values during translation and rotation.

■ **B:** The length of a vector $\vec{\mathbf{a}} = \{a_1, a_2, a_3\}$, i.e., the quantity $\sqrt{a_1^2 + a_2^2 + a_3^2}$, is a scalar invariant.

■ **C:** The scalar product of a vector with itself is a scalar invariant:
$\vec{\mathbf{a}}\vec{\mathbf{a}} = \vec{\mathbf{a}}^2 = |\vec{\mathbf{a}}|^2 \cos\varphi = |\vec{\mathbf{a}}^2|$, because $\varphi = 0$.

3.5.3.5 Cardan Angles

1. Definition of the Cardan angles

Every rotation of a coordinate system around an arbitrary axis through the origin can be described by three successive rotations around the coordinate axis. The angles of rotation α, β, γ are called the CARDAN-angles, if the rotations are performed in the following order (see schematically **Fig.3.176**):

1. The first rotation is around the z-axis by an angle γ, and results in the $x^{(1)}$, $y^{(1)}$, $z^{(1)}$ coordinate system with $z^{(1)} = z$.

2. The second rotation is made around the image of the y-axis by the first rotation, i.e. around the $y^{(1)}$ axis by an angle β, and results the $x^{(2)}$, $y^{(2)}$, $z^{(2)}$ coordinate system with $y^{(2)} = y^{(1)}$.

3. The third rotation is made around the image of the x-axis by the second rotation, i.e. around the $x^{(2)}$-axis by an angle α. It results the finally required x', y', z' coordinate system, $x' = x^{(3)} = x^{(2)}$, $y' = y^{(3)}$, $z' = z^{(3)}$.

Remark: There are different definitions of the Cardan-angles given in the literature.

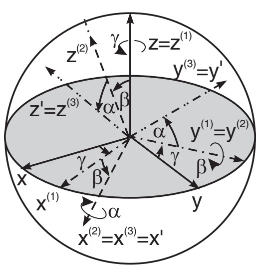


Figure 3.176

2. Calculation of the rotation matrix $\mathbf{D}_C$

It follows from the given order of the rotations (3.360a-d) that the rotation matrix $\mathbf{D} = \mathbf{D}_C$ is

$$\mathbf{D}_C = \mathbf{D}_x(\alpha)\mathbf{D}_y(\beta)\mathbf{D}_z(\gamma)\,. \tag{3.363a}$$

According to the Falk's schema (**Fig. 4.1, 4.2**, p. 273) holds

$$\begin{array}{c|c} & \mathbf{D}_z(\gamma) \\ \hline \mathbf{D}_y(\beta) & \mathbf{D}_y(\beta)\mathbf{D}_z(\gamma) \\ \hline \mathbf{D}_x(\alpha) & \mathbf{D}_C \end{array}\,,\ \text{i.e.} \tag{3.363b}$$

$$\begin{array}{ccc|ccc}
 & & & \cos\gamma & \sin\gamma & 0 \\
 & & & -\sin\gamma & \cos\gamma & 0 \\
 & & & 0 & 0 & 1 \\ \hline
\cos\beta & 0 & -\sin\beta & \cos\beta\cos\gamma & \cos\beta\sin\gamma & -\sin\beta \\
0 & 1 & 0 & -\sin\gamma & \cos\gamma & 0 \\
\sin\beta & 0 & \cos\beta & \cos\gamma\sin\beta & \sin\beta\sin\gamma & \cos\beta \\ \hline
1 & 0 & 0 & \cos\beta\cos\gamma & \cos\beta\sin\gamma & -\sin\beta \\
0 & \cos\alpha & \sin\alpha & -\cos\alpha\sin\gamma+\sin\alpha\cos\gamma\sin\beta & \cos\alpha\cos\gamma+\sin\alpha\sin\beta\sin\gamma & \sin\alpha\cos\beta \\
0 & -\sin\alpha & \cos\alpha & \sin\alpha\sin\gamma+\cos\alpha\cos\gamma\sin\beta & \cos\alpha\sin\beta\sin\gamma-\sin\alpha\cos\gamma & \cos\alpha\cos\beta
\end{array}$$

$$\mathbf{D}_C = \begin{pmatrix} \cos\beta\cos\gamma & \cos\beta\sin\gamma & -\sin\beta \\ -\cos\alpha\sin\gamma+\sin\alpha\cos\gamma\sin\beta & \cos\alpha\cos\gamma+\sin\alpha\sin\beta\sin\gamma & \sin\alpha\cos\beta \\ \sin\alpha\sin\gamma+\cos\alpha\cos\gamma\sin\beta & \cos\alpha\sin\beta\sin\gamma-\sin\alpha\cos\gamma & \cos\alpha\cos\beta \end{pmatrix}. \tag{3.363c}$$

3. Direction Cosines as Function of the Cardan angles

Since the rotation matrices $\mathbf{D}$ (in (3.362c), p. 213) and $\mathbf{D}_C$ (in (3.363c)) coincide, the direction cosines can be expressed as functions of the Cardan-angles:

$$\begin{array}{lll} l_1 = c_2c_3\,, & m_1 = c_2s_3\,, & n_1 = s_2\,; \\ l_2 = -c_1s_3 + s_1c_3s_2\,, & m_2 = c_1c_3 + s_1s_2s_3\,, & n_2 = s_1c_2\,; \\ l_3 = s_1s_3 + c_1c_3s_2\,, & m_3 = c_1s_2s_3 - s_1c_3\,, & n_3 = c_1c_3 \end{array} \tag{3.364a}$$

$$\text{with}\quad \begin{array}{lll} c_1 = \cos\alpha\,, & c_2 = \cos\beta\,, & c_3 = \cos\gamma\,, \\ s_1 = \sin\alpha\,, & s_2 = \sin\beta\,, & s_3 = \sin\gamma\,. \end{array} \tag{3.364b}$$

3.5.3.6 Euler's angles

1. Definition of the Euler angles

The position of the new coordinate system with respect to the old one can be uniquely determined by three angles, introduced by Euler **(Fig. 3.177)**.

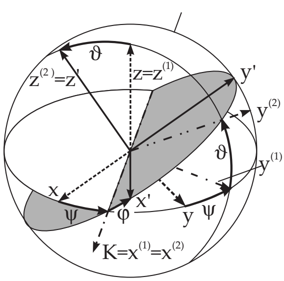


Figure 3.177

a) The **nutation angle** ϑ is the angle between the positive halves of the z- and z'-axis there is $0 \le \vartheta < \pi$.

b) The **precession angle** ψ is the angle between the positive direction of the x-axis and the intersection line K of the x, y plane and the x', y' plane. The positive direction of K is chosen depending on whether the z-axis, the z'-axis and K form a direction triplet with the same orientation as the coordinate axes (see 3.5.1.3, **2.**, p. 183). The angle ψ is measured from the x-axis to the direction of the y-axis; there is $0 \le \psi < \pi$.

c) The **rotation angle** φ is the angle between the positive halves of the x'-axis and the intersection line K; there is $0 \le \varphi < 2\pi$.

Remark: In the literature there are also other definitions in use for the Euler angles.

2. Calculation of the rotation matrix $\mathbf{D}_\mathrm{E}$

The transition from coordinate system x, y, z into the coordinate system x', y', z' (**Abb.3.177**) can be given by three rotations considering (3.360a-d)) as follows:

1. The first rotation is around the z-axis by angle ψ, and results the coordinate system $x^{(1)}, y^{(1)}, z^{(1)}$, where $z^{(1)} = z$:

$$\begin{pmatrix} x^{(1)} \\ y^{(1)} \\ z^{(1)} \end{pmatrix} = \begin{pmatrix} \cos\psi & \sin\psi & 0 \\ -\sin\psi & \cos\psi & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix}, \quad \text{i.e. } \underline{\mathbf{x}}^{(1)} = \mathbf{D}_z(\psi)\,\underline{\mathbf{x}}. \tag{3.365a}$$

The axis $x^{(1)}$ coincides with the intersection line K.

2. The second rotation is around the $x^{(1)}$-axis by angle ϑ, and results the coordinate system $x^{(2)}, y^{(2)}, z^{(2)}$, where $x^{(2)} = x^{(1)}$:

$$\begin{pmatrix} x^{(2)} \\ y^{(2)} \\ z^{(2)} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\vartheta & \sin\vartheta \\ 0 & -\sin\vartheta & \cos\vartheta \end{pmatrix} \begin{pmatrix} x^{(1)} \\ y^{(1)} \\ z^{(1)} \end{pmatrix}, \quad \text{i.e. } \underline{\mathbf{x}}^{(2)} = \mathbf{D}_x(\vartheta)\,\underline{\mathbf{x}}^{(1)}. \tag{3.365b}$$

3. The third rotation is around the $z^{(2)}$-axis by angle φ, and results the final coordinate system x', y', z', where $z' = z^{(2)}$:

$$\begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = \begin{pmatrix} \cos\varphi & \sin\varphi & 0 \\ -\sin\varphi & \cos\varphi & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x^{(2)} \\ y^{(2)} \\ z^{(2)} \end{pmatrix}, \quad \text{i.e. } \underline{\mathbf{x}}' = \mathbf{D}_z(\varphi)\,\underline{\mathbf{x}}^{(2)}. \tag{3.365c}$$

All together for $\mathbf{D} = \mathbf{D}_\mathrm{E}$ in (3.360a) holds:

$$\mathbf{D}_\mathrm{E} = \mathbf{D}_z(\varphi)\mathbf{D}_x(\vartheta)\mathbf{D}_z(\psi). \tag{3.366a}$$

Analogous to the calculation of the rotation matrix $\mathbf{D}_\mathrm{C}$ using Falk's schema, here in the form

	$\mathbf{D}_z(\psi)$
$\mathbf{D}_x(\vartheta)$	$\mathbf{D}_x(\vartheta)\mathbf{D}_z(\psi)$
$\mathbf{D}_z(\varphi)$	$\mathbf{D}_z(\varphi)\mathbf{D}_x(\vartheta)\mathbf{D}_z(\psi) = \mathbf{D}_\mathrm{E}$

one gets (3.366b)

$$\mathbf{D}_\mathrm{E} = \begin{pmatrix} \cos\varphi\cos\psi - \sin\varphi\cos\vartheta\sin\psi & \cos\varphi\sin\psi + \sin\varphi\cos\vartheta\cos\psi & \sin\varphi\sin\vartheta \\ -\sin\varphi\cos\psi - \cos\varphi\cos\vartheta\sin\psi & -\sin\varphi\sin\psi + \cos\varphi\cos\vartheta\cos\psi & \cos\varphi\sin\vartheta \\ \sin\vartheta\sin\psi & -\sin\vartheta\cos\psi & \cos\vartheta \end{pmatrix} \tag{3.366c}$$

3. Direction Cosines as Function of the Euler angles

Because of the identity of the rotation matrices $\mathbf{D}$ (see (3.362c), p. 213) and $\mathbf{D}_\mathrm{E}$ (see (3.366c) the following formulas for the direction cosines as functions of the Euler angles are valid:

$$\begin{aligned} l_1 &= c_2c_3 - c_1s_2s_3, & m_1 &= s_2c_3 + c_1c_2s_3, & n_1 &= s_1s_3; \\ l_2 &= -c_2s_3 - c_1s_2c_3, & m_2 &= -s_2s_3 + c_1c_2c_3, & n_2 &= s_1c_3; \\ l_3 &= s_1s_2, & m_3 &= -s_1c_2, & n_3 &= c_1 \end{aligned} \tag{3.367a}$$

$$\text{with } \begin{matrix} c_1 = \cos\vartheta, & c_2 = \cos\psi, & c_3 = \cos\varphi, \\ s_1 = \sin\vartheta, & s_2 = \sin\psi, & s_3 = \sin\varphi. \end{matrix} \tag{3.367b}$$

3.5.3.7 Special Quantities in Space

1. Coordinates of the Center of Mass

The coordinates of the *center of mass* $P(\bar{x}, \bar{y}, \bar{z})$ (often called incorrectly the *center of gravity*) of a system of n material points $P_i(x_i, y_i, z_i)$ with mass m_i are calculated by the following formulas, where the sum index i changes from 1 to n:

$$\bar{x} = \frac{\sum m_i x_i}{\sum m_i}, \quad \bar{y} = \frac{\sum m_i y_i}{\sum m_i}, \quad \bar{z} = \frac{\sum m_i z_i}{\sum m_i}. \tag{3.368}$$

2. Division of a Segment

The coordinates of the point $P(x, y, z)$ dividing a segment between the points $P_1\,(x_1, y_1, z_1)$ and $P_2\,(x_2, y_2, z_2)$ in a given ratio

$$\frac{\overline{P_1P}}{\overline{PP_2}} = \frac{m}{n} = \lambda \qquad (3.369a)$$

are given by the formulas

$$x = \frac{nx_1 + mx_2}{n+m} = \frac{x_1 + \lambda x_2}{1+\lambda}, \quad (3.369b) \qquad y = \frac{ny_1 + my_2}{n+m} = \frac{y_1 + \lambda y_2}{1+\lambda}, \qquad (3.369c)$$

$$z = \frac{nz_1 + mz_2}{n+m} = \frac{z_1 + \lambda z_2}{1+\lambda}. \qquad (3.369d)$$

The *midpoint* of the segment is given by

$$x_m = \frac{x_1 + x_2}{2}, \quad y_m = \frac{y_1 + y_2}{2}, \quad z_m = \frac{z_1 + z_2}{2}. \qquad (3.369e)$$

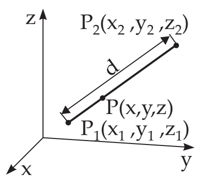


Figure 3.178

3. Distance Between Two Points

The distance between the points $P_1\,(x_1, y_1, z_1)$ and $P_2\,(x_2, y_2, z_2)$ in **Fig. 3.178** is

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}. \qquad (3.370a)$$

The direction cosines of the segment between the points can be calculated by the formulas

$$\cos\alpha = \frac{x_2 - x_1}{d}, \quad \cos\beta = \frac{y_2 - y_1}{d}, \quad \cos\gamma = \frac{z_2 - z_1}{d}. \qquad (3.370b)$$

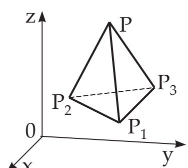


Figure 3.179

4. System of Four Points

Four points $P(x, y, z)$, $P_1\,(x_1, y_1, z_1)$, $P_2\,(x_2, y_2, z_2)$ and $P_3\,(x_3, y_3, z_3)$ can form a tetrahedron **(Fig. 3.179)** or they are in a plane. The volume of a tetrahedron can be calculated by the formula

$$V = \frac{1}{6}\begin{vmatrix} x & y & z & 1 \\ x_1 & y_1 & z_1 & 1 \\ x_2 & y_2 & z_2 & 1 \\ x_3 & y_3 & z_3 & 1 \end{vmatrix} = \frac{1}{6}\begin{vmatrix} x - x_1 & y - y_1 & z - z_1 \\ x - x_2 & y - y_2 & z - z_2 \\ x - x_3 & y - y_3 & z - z_3 \end{vmatrix}, \qquad (3.371)$$

where it has a positive value $V > 0$ if the orientation of the three vectors $\overrightarrow{PP_1}$, $\overrightarrow{PP_2}$, $\overrightarrow{PP_3}$ is the same as the coordinate axes (see 3.5.1.3, **2.**, p. 183). Otherwise it is negative.
The four points are in the same plane if and only if

$$\begin{vmatrix} x & y & z & 1 \\ x_1 & y_1 & z_1 & 1 \\ x_2 & y_2 & z_2 & 1 \\ x_3 & y_3 & z_3 & 1 \end{vmatrix} = 0 \qquad \text{holds.} \qquad (3.372)$$

3.5.3.8 Equation of a Surface

Often an equation

$$F(x, y, z) = 0 \qquad (3.373)$$

corresponds to a surface with the property that the coordinates of every of its points P satisfy this equation. Conversely, every point whose coordinates satisfy the equation is a point of this surface. The

equation (3.373) is called the equation of this surface. If there is no real point in the space satisfying equation (3.373), then there is no real surface.

1. The Equation of a Cylindrical Surface (see 3.3.4, p. 156) whose generating lines are parallel to the x-axis contains no x coordinate: $F(y,z)=0$. Similarly, the equations of the cylindrical surfaces with generating lines parallel to the y- or to the z-axes contain no y or z coordinates: $F(x,z)=0$ or $F(x,y)=0$ resp. The equation $F(x,y)=0$ describes the intersection curve between the cylinder and the x,y plane. If the direction cosines, or the proportional quantities l, m, n of the generating line of a cylinder are given, then the equation has the form

$$F(nx-lz,ny-mz)=0. \tag{3.374}$$

2. The Equation of a Rotationally Symmetric Surface, i.e., a surface which is created by the rotation of a curve $z=f(x)$ given in the x,z plane around the z-axis **(Fig. 3.180)**, will have the form

$$z=f\left(\sqrt{x^2+y^2}\right). \tag{3.375}$$

The equations of rotationally symmetric surfaces can be obtained similarly in the case of other variables as well.

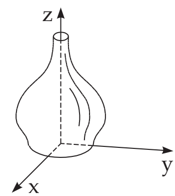


Figure 3.180

■ The equation of a *conical surface*, whose vertex is at the origin (see 3.3.4, p. 157), has the form $F(x,y,z)=0$, where F is a homogeneous function of the coordinates (see 2.18.2.5, **4.**, p. 122), e.g. $F(x,y,z)=z-\sqrt{x^2+y^2}=0$ with the degree 1 of homogeneity, i.e. $F(\lambda x,\lambda y,\lambda z)=\lambda F(x,y,z)$.

3.5.3.9 Equation of a Space Curve

A space curve can be defined by three parametric equations

$$x=\varphi_1(t)\,,\quad y=\varphi_2(t)\,,\quad z=\varphi_3(t)\,. \tag{3.376}$$

To every value of the parameter t, which does not necessarily have a geometrical meaning, there corresponds a point of the curve.
Another method to define a space curve is the determination by two equations

$$F_1(x,y,z)=0\,,\quad F_2(x,y,z)=0\,. \tag{3.377}$$

Both define a surface. The space curve contains all points whose coordinates satisfy both equations, i.e., the space curve is the intersection curve of the given surfaces. In general, every equation in the form

$$F_1+\lambda F_2=0 \tag{3.378}$$

for arbitrary λ defines a surface which goes through the considered curve, so it can substitute any of the equations (3.377).

3.5.3.10 Line and Plane in Space

1. Equations of the Plane

Every equation linear in the coordinates defines a plane, and conversely every plane has an equation of first degree.

1. General Equation of the Plane

a) with coordinates: $Ax+By+Cz+D=0,$ (3.379a)

b) in vector form: $\vec{\mathbf{r}}\vec{\mathbf{N}}+D=0,$* (3.379b)

*For the scalar product of two vectors see 3.5.1.5, p. 184 and in affine coordinates see 3.5.1.6, **5.**, p. 186; for the vector equation of the plane see 3.5.1.7, p. 189.

where the vector $\vec{\mathbf{N}}(A, B, C)$ is perpendicular to the plane. In **(Fig. 3.181)** the intercepts a, b, and c are shown. The vector $\vec{\mathbf{N}}$ is called the *normal vector of the plane*. Its direction cosines are

$$\cos\alpha = \frac{A}{\sqrt{A^2+B^2+C^2}}, \quad \cos\beta = \frac{B}{\sqrt{A^2+B^2+C^2}}, \quad \cos\gamma = \frac{C}{\sqrt{A^2+B^2+C^2}}. \tag{3.379c}$$

If $D = 0$ holds, the plane goes through the origin; for $A = 0$, or $B = 0$, or $C = 0$ the plane is parallel to the x-axis, the y-axis, or the z-axis, respectively. If $A = B = 0$, or $A = C = 0$, or $B = C = 0$, then the plane is parallel to the x, y plane, the x, z plane or the y, z plane, respectively.

2. Hessian Normal Form of the Equation of the Plane

a) with coordinates: $x\cos\alpha + y\cos\beta + z\cos\gamma - p = 0,$ (3.380a)

b) in vector form: $\vec{\mathbf{r}}\vec{\mathbf{N}}^{0} - p = 0,$ (3.380b)

where $\vec{\mathbf{N}}^{0}$ is the unit normal vector of the plane and p is the distance of the plane from the origin. The Hessian normal form arises from the general equation (3.379a) by multiplying by the normalizing factor

$$\pm\mu = \frac{1}{N} = \frac{1}{\sqrt{A^2+B^2+C^2}} \quad \text{with } N = |\vec{\mathbf{N}}|. \tag{3.380c}$$

Here the sign of μ must be chosen opposite to that of D.

3. Intercept Form of the Equation of the Plane With the segments a, b, c, considering them with signs depending on where the plane intersects the coordinate axes **(Fig. 3.181)** holds:

$$\frac{x}{a} + \frac{y}{b} + \frac{z}{c} = 1. \tag{3.381}$$

4. Equation of the Plane Through Three Points If the points are $P_1(x_1, y_1, z_1)$, $P_2(x_2, y_2, z_2)$, $P_3(x_3, y_3, z_3)$, then it holds:

a) with coordinates:
$$\begin{vmatrix} x - x_1 & y - y_1 & z - z_1 \\ x_2 - x_1 & y_2 - y_1 & z_2 - z_1 \\ x_3 - x_1 & y_3 - y_1 & z_3 - z_1 \end{vmatrix} = 0, \tag{3.382a}$$

b) in vector form: $(\vec{\mathbf{r}} - \vec{\mathbf{r}}_1)(\vec{\mathbf{r}} - \vec{\mathbf{r}}_2)(\vec{\mathbf{r}} - \vec{\mathbf{r}}_3) = 0^{\dagger}.$ (3.382b)

5. Equation of a Plane Through Two Points and Parallel to a Line

The equation of the plane passing through the two points $P_1(x_1, y_1, z_1)$, $P_2(x_2, y_2, z_2)$ and being parallel to the line with direction vector $\vec{\mathbf{R}}(l, m, n)$ is the following

a) with coordinates:
$$\begin{vmatrix} x - x_1 & y - y_1 & z - z_1 \\ x_2 - x_1 & y_2 - y_1 & z_2 - z_1 \\ l & m & n \end{vmatrix} = 0, \tag{3.383a}$$

b) in vector form: $(\vec{\mathbf{r}} - \vec{\mathbf{r}}_1)(\vec{\mathbf{r}} - \vec{\mathbf{r}}_2)\vec{\mathbf{R}} = 0^{\dagger}.$ (3.383b)

6. Equation of a Plane Through a Point and Parallel to Two Lines

If the direction vectors of the lines are $\vec{\mathbf{R}}_1(l_1, m_1, n_1)$ and $\vec{\mathbf{R}}_2(l_2, m_2, n_2)$, then it holds:

a) with coordinates:
$$\begin{vmatrix} x - x_1 & y - y_1 & z - z_1 \\ l_1 & m_1 & n_1 \\ l_2 & m_2 & n_2 \end{vmatrix} = 0, \tag{3.384a}$$

b) in vector form: $(\vec{\mathbf{r}} - \vec{\mathbf{r}}_1)\vec{\mathbf{R}}_1\vec{\mathbf{R}}_2 = 0^{\dagger}.$ (3.384b)

7. Equation of a Plane Through a Point and Perpendicular to a Line

If the point is $P_1(x_1, y_1, z_1)$, and the direction vector of the line is $\vec{\mathbf{N}}(A, B, C)$, then it holds:

a) with coordinates: $A(x - x_1) + B(y - y_1) + C(z - z_1) = 0,$ (3.385a)

†For the mixed product of three vectors see 3.5.1.6, **2.**, p. 185

b) in vector form: $(\vec{\mathbf{r}} - \vec{\mathbf{r}}_1)\,\vec{\mathbf{N}} = 0\,.$* (3.385b)

8. Distance of a Point from a Plane Substituting the coordinates of the point $P(a, b, c)$ in the Hessian normal form of the equation of the plane (3.380a)

$$x\cos\alpha + y\cos\beta + z\cos\gamma - p = 0, \tag{3.386a}$$

results in the distance with sign

$$\delta = a\cos\alpha + b\cos\beta + c\cos\delta - p, \tag{3.386b}$$

where $\delta > 0$, if P and the origin are on different sides of the plane; in the opposite case $\delta < 0$ holds.

9. Equation of a Plane Through the Intersection Line of Two Planes The equation of a plane which goes through the intersection line of the planes given by the equations $A_1x+B_1y+C_1z+D_1 = 0$ and $A_2x + B_2y + C_2z + D_2 = 0$ is

a) with coordinates: $A_1x + B_1y + C_1z + D_1 + \lambda\,(A_2x + B_2y + C_2z + D_2) = 0.$ (3.387a)

b) in vector form: $\vec{\mathbf{r}}\,\vec{\mathbf{N}}_1 + D_1 + \lambda(\vec{\mathbf{r}}\,\vec{\mathbf{N}}_2 + D_2) = 0.$ (3.387b)

Here λ is a real parameter, so (3.387a) and (3.387b) define a pencil of planes. **Fig. 3.182** shows the case with three planes. If λ takes all the values between $-\infty$ and $+\infty$ in (3.387a) and (3.387b), so this yields all the planes from the pencil. For $\lambda = \pm 1$ it results in the equations of the planes bisecting the angle between the given planes if their equations are in normal form.

2. Two and More Planes in Space

1. Angle between Two Planes, General Case: The angle φ between two planes given by the equations $A_1x+B_1y+C_1z+D_1 = 0$ and $A_2x+B_2y+C_2z+D_2 = 0$ can be calculated by the formula

$$\cos\varphi = \frac{A_1A_2 + B_1B_2 + C_1C_2}{\sqrt{(A_1^2 + B_1^2 + C_1^2)\,(A_2^2 + B_2^2 + C_2^2)}}. \tag{3.388a}$$

If the planes are given by vector equations $\vec{\mathbf{r}}\,\vec{\mathbf{N}}_1 + D_1 = 0$ and $\vec{\mathbf{r}}\,\vec{\mathbf{N}}_2 + D_2 = 0$, then

$$\cos\varphi = \frac{\vec{\mathbf{N}}_1\,\vec{\mathbf{N}}_2}{N_1N_2} \quad \text{with} \quad N_1 = |\vec{\mathbf{N}}_1| \text{ and } N_2 = |\vec{\mathbf{N}}_2|. \tag{3.388b}$$

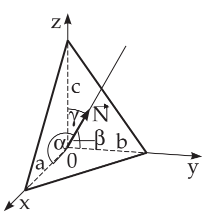


Figure 3.181

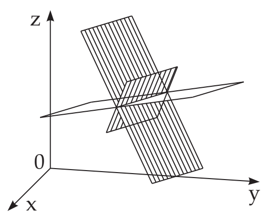


Figure 3.182

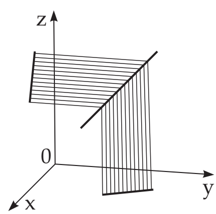


Figure 3.183

2. Intersection Point of Three Planes: The coordinates of the intersection point of three planes given by the three equations $A_1x+B_1y+C_1z+D_1 = 0$, $A_2x+B_2y+C_2z+D_2 = 0$, and $A_3x+B_3y+C_3z+D_3 = 0$, are calculated by the formulas

$$\bar{x} = \frac{-\Delta x}{\Delta}, \quad \bar{y} = \frac{-\Delta y}{\Delta}, \quad \bar{z} = \frac{-\Delta z}{\Delta} \quad \text{with} \tag{3.389a}$$

*For the scalar product of two vectors see 3.5.1.5, p. 184 and in affine coordinates see 3.5.1.6, p. 186; for the equation of the plane in vector form see 3.5.1.6, p. 189.

$$\Delta = \begin{vmatrix} A_1 & B_1 & C_1 \\ A_2 & B_2 & C_2 \\ A_3 & B_3 & C_3 \end{vmatrix}, \quad \Delta_x = \begin{vmatrix} D_1 & B_1 & C_1 \\ D_2 & B_2 & C_2 \\ D_3 & B_3 & C_3 \end{vmatrix}, \quad \Delta_y = \begin{vmatrix} A_1 & D_1 & C_1 \\ A_2 & D_2 & C_2 \\ A_3 & D_3 & C_3 \end{vmatrix}, \quad \Delta_z = \begin{vmatrix} A_1 & B_1 & D_1 \\ A_2 & B_2 & D_2 \\ A_3 & B_3 & D_3 \end{vmatrix}. \quad (3.389b)$$

Three planes intersect each other at one point if $\Delta \neq 0$ holds. If $\Delta = 0$ holds and at least one subdeterminant of second order is non-zero, then the planes are parallel to a line; if every subdeterminant of second order is zero, then the planes have a common line.

3. Conditions for Parallelism and Orthogonality of Planes:

a) Conditions for Parallelism: Two planes are parallel if

$$\frac{A_1}{A_2} = \frac{B_1}{B_2} = \frac{C_1}{C_2} \quad \text{or} \quad \vec{\mathbf{N}}_1 \times \vec{\mathbf{N}}_2 = \vec{\mathbf{0}} \quad \text{holds.} \quad (3.390)$$

b) Conditions for Orthogonality: Two planes are perpendicular to each other if

$$A_1A_2 + B_1B_2 + C_1C_2 = 0 \quad \text{or} \quad \vec{\mathbf{N}}_1\vec{\mathbf{N}}_2 = 0 \quad \text{holds.} \quad (3.391)$$

4. Intersection Point of Four Planes: Four planes given by the equations $A_1x+B_1y+C_1z+D_1 = 0$, $A_2x + B_2y + C_2z + D_2 = 0$, $A_3x + B_3y + C_3z + D_3 = 0$, and $A_4x + B_4y + C_4z + D_4 = 0$ have a common point only if for the determinant

$$\delta = \begin{vmatrix} A_1 & B_1 & C_1 & D_1 \\ A_2 & B_2 & C_2 & D_2 \\ A_3 & B_3 & C_3 & D_3 \\ A_4 & B_4 & C_4 & D_4 \end{vmatrix} = 0 \quad (3.392)$$

holds. In this case the common point is determined from three equations, the fourth equation is superfluous; it is a consequence of the others.

5. Distance Between Two Parallel Planes: If two planes are parallel, and they are given by the equations

$$Ax + By + Cz + D_1 = 0 \quad \text{and} \quad Ax + By + Cz + D_2 = 0, \quad (3.393)$$

then their distance is

$$d = \frac{|D_1 - D_2|}{\sqrt{A^2 + B^2 + C^2}}. \quad (3.394)$$

3.5.3.11 Lines in Space

1. Equations of a Line

1. Equation of a Line in Space, General Case Because a line in space can be defined as the intersection of two planes, it can be represented by a system of two linear equations.

a) In component form:

$$A_1x + B_1y + C_1z + D_1 = 0, \quad A_2x + B_2y + C_2z + D_2 = 0. \quad (3.395a)$$

b) In vector form:

$$\vec{\mathbf{r}}\vec{\mathbf{N}}_1 + D_1 = 0, \quad \vec{\mathbf{r}}\vec{\mathbf{N}}_2 + D_2 = 0. \quad (3.395b)$$

2. Equation of a Line in Two Projecting Planes

The two equations $y = kx + a, \quad z = hx + b$ (3.396)

define a plane each, and these planes go through the line and are perpendicular to the x, y and the x, z planes resp. **(Fig. 3.183)**. They are called *projecting planes*. This representation cannot be used for lines parallel to the y, z plane, so in this case other projections to other coordinate planes are considered.

3. Equation of a Line Through a Point Parallel to a Direction Vector

The equation (or the system of equations) of a line passing through a point $P_1(x_1, y_1, z_1)$ parallel to a direction vector $\vec{\mathbf{R}}(l, m, n)$ **(Fig. 3.184)** has the form

a) in component representation and in vector form:

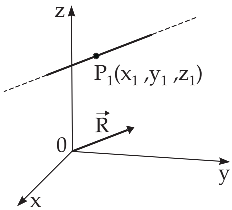


Figure 3.184

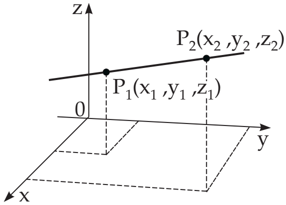


Figure 3.185

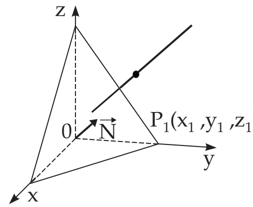


Figure 3.186

$$\frac{x-x_1}{l}=\frac{y-y_1}{m}=\frac{z-z_1}{n}, \quad (3.397a) \qquad (\vec{\mathbf{r}}-\vec{\mathbf{r}}_1)\times\vec{\mathbf{R}}=\vec{\mathbf{0}}, \quad (3.397b)$$

b) in parametric form and vector form:

$$x=x_1+l\,t,\quad y=y_1+m\,t,\quad z=z_1+n\,t; \quad (3.397c) \qquad \vec{\mathbf{r}}=\vec{\mathbf{r}}_1+\vec{\mathbf{R}}t, \quad (3.397d)$$

where the numbers x_1, y_1, z_1 are chosen such that the equations in (3.395a) are satisfied. The representation (3.397a) follows from (3.395a) with

$$l=\begin{vmatrix} B_1 & C_1 \\ B_2 & C_2 \end{vmatrix},\quad m=\begin{vmatrix} C_1 & A_1 \\ C_2 & A_2 \end{vmatrix},\quad n=\begin{vmatrix} A_1 & B_1 \\ A_2 & B_2 \end{vmatrix}, \quad (3.398a)$$

or in vector form $\vec{\mathbf{R}}=\vec{\mathbf{N}}_1\times\vec{\mathbf{N}}_2$. (3.398b)

4. Equation of a Line Through Two Points The equation of a line through two points $P_1\,(x_1,y_1,z_1)$ and $P_2\,(x_2,y_2,z_2)$ **(Fig. 3.185)** is
in component form and in vector form:

a) $$\frac{x-x_1}{x_2-x_1}=\frac{y-y_1}{y_2-y_1}=\frac{z-z_1}{z_2-z_1}, \quad (3.399a)$$ **b)** $$(\vec{\mathbf{r}}-\vec{\mathbf{r}}_1)\times(\vec{\mathbf{r}}-\vec{\mathbf{r}}_2)=\vec{\mathbf{0}}^*. \quad (3.399b)$$

If for instance $x_1=x_2$, the equations in component form are $x=x_2$, $\dfrac{y-y_1}{y_2-y_1}=\dfrac{z-z_1}{z_2-z_1}$. If $x_1=x_2$ and $y_1=y_2$ are both valid, the equations in component form are $x=x_2$, $y=y_2$.

5. Equation of a Line Through a Point and Perpendicular to a Plane The equation of a line passing through the point $P_1(x_1,y_1,z_1)$ and being perpendicular to a plane given by the equation $Ax+By+Cz+D=0$ or by $\vec{\mathbf{r}}\vec{\mathbf{N}}+D=0$ **(Fig. 3.186)** is
in component form and in vector form:

a) $$\frac{x-x_1}{A}=\frac{y-y_1}{B}=\frac{z-z_1}{C}, \quad (3.400a)$$ **b)** $$(\vec{\mathbf{r}}-\vec{\mathbf{r}}_1)\times\vec{\mathbf{N}}=\vec{\mathbf{0}}. \quad (3.400b)$$

If for instance $A=0$ holds, the equations in component form have a similar form as in the previous case.

2. Distance of a Point from a Line Given in Component Form

For the distance d of the point $M(a,b,c)$ from a line given in the form (3.397a) holds:

$$d^2=\frac{[(a-x_1)\,m-(b-y_1)\,l]^2+[(b-y_1)\,n-(c-z_1)\,m]^2+[(c-z_1)\,l-(a-x_1)\,n]^2}{l^2+m^2+n^2}. \quad (3.401)$$

*For the product of vectors see 3.5.1.5, p. 184

3. Smallest Distance Between Two Lines Given in Component Form
If the lines are given in the form (3.397a), their distance is

$$d = \frac{\pm \begin{vmatrix} x_1 - x_2 & y_1 - y_2 & z_1 - z_2 \\ l_1 & m_1 & n_1 \\ l_2 & m_2 & n_2 \end{vmatrix}}{\sqrt{\begin{vmatrix} l_1 & m_1 \\ l_2 & m_2 \end{vmatrix}^2 + \begin{vmatrix} m_1 & n_1 \\ m_2 & n_2 \end{vmatrix}^2 + \begin{vmatrix} n_1 & l_1 \\ n_2 & l_2 \end{vmatrix}^2}} . \tag{3.402}$$

If the determinant in the numerator is equal to zero, the lines intersect each other.

3.5.3.12 Intersection Points and Angles of Lines and Planes in Space

1. Intersection Points of Lines and Planes

1. Equation of the Line in Component Form The intersection point of a plane given by the equation
$Ax + By + Cz + D = 0$, and a line given by $\frac{x - x_1}{l} = \frac{y - y_1}{m} = \frac{z - z_1}{n}$ has the coordinates

$$\bar{x} = x_1 - l\rho, \quad \bar{y} = y_1 - m\rho, \quad \bar{z} = z_1 - n\rho \qquad \text{with} \tag{3.403a}$$

$$\rho = \frac{Ax_1 + By_1 + Cz_1 + D}{A\,l + B\,m + C\,n} . \tag{3.403b}$$

If $A\,l + B\,m + C\,n = 0$ holds, then the line is parallel to the plane. If $Ax_1 + By_1 + Cz_1 + D = 0$ is also valid, then the line lies in the plane.
2. Equation of the Line in Two Projecting Planes The intersection point of a plane given by the equation $Ax + By + Cz + D = 0$, and a line given by $y = kx + a$, and $z = hx + b$ has the coordinates

$$\bar{x} = -\frac{B\,a + C\,b + D}{A + B\,k + C\,h}, \quad \bar{y} = k\bar{x} + a, \quad \bar{z} = h\bar{x} + b. \tag{3.404}$$

If $A + B\,k + C\,h = 0$ holds, then the line is parallel to the plane. If $Ba + Cb + D = 0$ is also valid, then the line lies in the plane.
3. Intersection Point of Two Lines If the lines are given by $y = k_1x + a_1,\ z = h_1x + b_1$ and $y = k_2x + a_2,\ z = h_2x + b_2$, then the coordinates of the intersection point, if any exists, are:

$$\bar{x} = \frac{a_2 - a_1}{k_1 - k_2} = \frac{b_2 - b_1}{h_1 - h_2}, \quad \bar{y} = \frac{k_1a_2 - k_2a_1}{k_1 - k_2}, \quad \bar{z} = \frac{h_1b_2 - h_2b_1}{h_1 - h_2} . \tag{3.405a}$$

The intersection point exists only if

$$(a_1 - a_2)(h_1 - h_2) = (b_1 - b_2)(k_1 - k_2). \tag{3.405b}$$

Otherwise the lines do not intersect each other.

2. Angles between Planes and Lines

1. Angle between Two Lines

a) General Case: If the lines are given by the equations $\frac{x - x_1}{l_1} = \frac{y - y_1}{m_1} = \frac{z - z_1}{n_1}$ and $\frac{x - x_2}{l_2} = \frac{y - y_2}{m_2} = \frac{z - z_2}{n_2}$ or in vector form by $(\vec{\mathbf{r}} - \vec{\mathbf{r}}_1) \times \vec{\mathbf{R}}_1 = \vec{\mathbf{0}}$ and $(\vec{\mathbf{r}} - \vec{\mathbf{r}}_2) \times \vec{\mathbf{R}}_2 = \vec{\mathbf{0}}$, then for the angle of intersection between them

$$\cos\varphi = \frac{l_1l_2 + m_1m_2 + n_1n_2}{\sqrt{(l_1^2 + m_1^2 + n_1^2)\,(l_2^2 + m_2^2 + n_2^2)}} \qquad \text{or} \tag{3.406a}$$

$$\cos\varphi = \frac{\vec{\mathbf{R}}_1\vec{\mathbf{R}}_2}{R_1R_2} \quad \text{with } R_1 = |\vec{\mathbf{R}}_1| \text{ and } R_2 = |\vec{\mathbf{R}}_2|. \tag{3.406b}$$

b) Conditions of Parallelism: Two lines are parallel if

$$\frac{l_1}{l_2} = \frac{m_1}{m_2} = \frac{n_1}{n_2} \qquad \text{or} \qquad \vec{\mathbf{R}}_1 \times \vec{\mathbf{R}}_2 = \vec{\mathbf{0}}. \tag{3.407}$$

c) Conditions of Orthogonality: Two lines are perpendicular to each other if

$$l_1 l_2 + m_1 m_2 + n_1 n_2 = 0 \qquad \text{or} \qquad \vec{\mathbf{R}}_1 \vec{\mathbf{R}}_2 = 0. \tag{3.408}$$

2. Angle Between a Line and a Plane

a) If the line and the plane are given by the equations $\frac{x-x_1}{l} = \frac{y-y_1}{m} = \frac{z-z_1}{n}$ and $Ax+By+Cz+D=0$ or in vector form by $(\vec{\mathbf{r}} - \vec{\mathbf{r}}_1) \times \vec{\mathbf{R}} = \vec{\mathbf{0}}$ and $\vec{\mathbf{r}}\,\vec{\mathbf{N}} + D = 0$, one gets the angle φ by the formulas

$$\sin\varphi = \frac{Al+Bm+Cn}{\sqrt{(A^2+B^2+C^2)\,(l^2+m^2+n^2)}} \quad \text{or} \tag{3.409a}$$

$$\sin\varphi = \frac{\vec{\mathbf{R}}\,\vec{\mathbf{N}}}{RN} \quad \text{with } R = |\vec{\mathbf{R}}| \text{ and } N = |\vec{\mathbf{N}}|. \tag{3.409b}$$

b) Conditions of Parallelism: A line and a plane are parallel if

$$A\,l + B\,m + C\,n = 0 \qquad \text{or} \qquad \vec{\mathbf{R}}\,\vec{\mathbf{N}} = 0. \tag{3.410}$$

c) Conditions of Orthogonality: A line and a plane are orthogonal if

$$\frac{A}{l} = \frac{B}{m} = \frac{C}{n} \qquad \text{or} \qquad \vec{\mathbf{R}} \times \vec{\mathbf{N}} = \vec{\mathbf{0}}. \tag{3.411}$$

3.5.3.13 Surfaces of Second Order, Equations in Normal Form

1. Central Surfaces

The following equations, which are also called the normal form of the equations of surfaces of second order, can be derived from the general equations of surfaces of second order (see 3.5.3.14, **1.**, p. 228) by putting the center at the origin. Here the center is the midpoint of the chords passing through it. The coordinate axes are the symmetry axes of the surfaces, so the coordinate planes are also the planes of symmetry.

2. Ellipsoid

With the semi-axes a, b, c **(Fig. 3.187)** the equation of an ellipsoid is

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1. \tag{3.412}$$

The following special cases are to be distinguished:

a) Compressed Ellipsoid of Revolution (Lens Form): $a = b > c$ **(Fig. 3.188)**.

b) Stretched Ellipsoid of Revolution (Cigar Form): $a = b < c$ **(Fig. 3.189)**.

c) Sphere: $a = b = c$ so that $x^2 + y^2 + z^2 = a^2$ is valid.

The two forms of the ellipsoid of revolution arise by rotating an ellipse in the x, z plane with axes a and c around the z-axis, and one gets a sphere if rotating a circle around any axis. If a plane goes through an ellipsoid, the intersection figure is an ellipse; in a special case it is a circle. The volume of the ellipsoid is

$$V = \frac{4\pi abc}{3}. \tag{3.413}$$

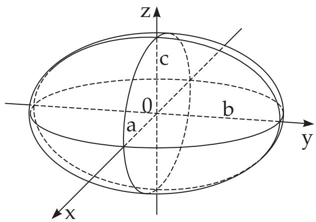


Figure 3.187

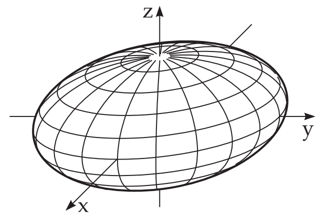


Figure 3.188

3. Hyperboloid

a) Hyperboloid of One Sheet (Fig. 3.190): With a and b as real and c as imaginary semi-axes the equation is

$$\frac{x^2}{a^2}+\frac{y^2}{b^2}-\frac{z^2}{c^2}=1 \quad \text{(for generator lines see p. 226).} \tag{3.414}$$

b) Hyperboloid of Two Sheets (Fig. 3.191): With c as real and a, b as imaginary semi-axes the equation is

$$\frac{x^2}{a^2}+\frac{y^2}{b^2}-\frac{z^2}{c^2}=-1\,. \tag{3.415}$$

Intersecting it by a plane parallel to the z-axis results in a hyperbola in the case of both types of hyperboloids. In the case of a hyperboloid of one sheet the intersection can also be two lines intersecting each other. The intersection figures parallel to the x, y plane are ellipses in both cases.
For $a = b$ the hyperboloid can be represented by rotation of a hyperbola with semi-axes a and c around the axis $2c$. This is imaginary in the case of a hyperboloid of one sheet, and real in the case of that of two sheets.

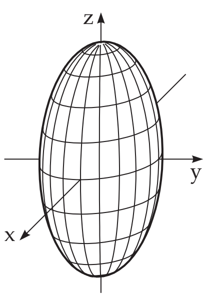


Figure 3.189

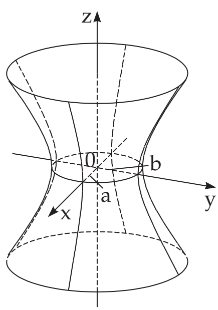


Figure 3.190

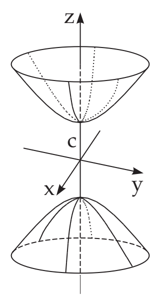


Figure 3.191

4. Cone (Fig. 3.192)

If the vertex is at the origin, the equation is

$$\frac{x^2}{a^2}+\frac{y^2}{b^2}-\frac{z^2}{c^2}=0\,. \tag{3.416}$$

As a direction curve can be considerd an ellipse with semi-axes a and b, whose plane is perpendicular to the z-axis at a distance c from the origin. The cone in this representation can be considered as the

asymptotic cone of the surfaces

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} - \frac{z^2}{c^2} = \pm 1, \tag{3.417}$$

whose generator lines approach infinitely closely both hyperboloids at infinity **(Fig. 3.193)**. For $a = b$ there is a right circular cone (see 3.3.4, **9.**, p. 157).

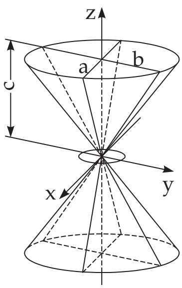


Figure 3.192

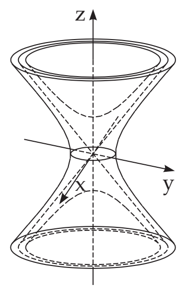


Figure 3.193

5. Paraboloid

Since a paraboloid has no center, in the following section it is supposed that the vertex is at the origin, the z-axis is its symmetry axis, and the x, z plane and the y, z plane are symmetry planes.

a) Elliptic Paraboloid (Fig. 3.194):

$$z = \frac{x^2}{a^2} + \frac{y^2}{b^2}\,. \tag{3.418}$$

The plane sections parallel to the z-axis result in parabolas as intersection figures; those parallel to the x, y plane result in ellipses.

The volume of a paraboloid which is cut by a plane perpendicular to the z-axis at a distance h from the origin is given as (see also **Fig.3.194**)

$$V = \frac{1}{2}\pi \bar{a}\bar{b}h\,, \tag{3.419}$$

The parameters $\overline{a} = a\sqrt{h}$ and $\overline{b} = b\sqrt{h}$ are the half axis of the intersecting ellipse at height h. So, (3.419) yields the half of the volume of an elliptic cylinder with the same upper surface and altitude (compare with (3.328a), p. 201).

b) Paraboloid of Revolution: For $a = b$ one gets a paraboloid of revolution. It can be generated by rotating the $z = x^2/a^2$ parabola of the x, z plane around the z-axis.

c) Hyperbolic Paraboloid (Fig. 3.195):

$$z = \frac{x^2}{a^2} - \frac{y^2}{b^2}\,. \tag{3.420}$$

The intersection figures parallel to the y, z plane or to the x, z plane are parabolas; parallel to the x, y plane they are hyperbolas or two intersecting lines.

6. Rectilinear Generators of a Ruled Surface

These are straight lines lying completely in this surface. Examples are the generators of the surfaces of the cone and cylinder.

a) Hyperboloid of One Sheet (Fig. 3.196):

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} - \frac{z^2}{c^2} = 1. \tag{3.421}$$

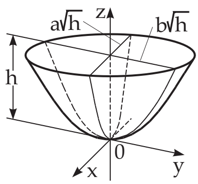


Figure 3.194

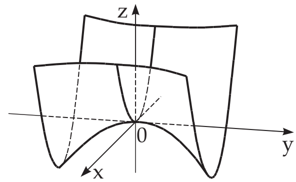


Figure 3.195

The hyperboloid of one sheet has two families of rectilinear generators with equations

$$\frac{x}{a} + \frac{z}{c} = u\left(1 + \frac{y}{b}\right), \quad u\left(\frac{x}{a} - \frac{z}{c}\right) = 1 - \frac{y}{b}; \tag{3.422a}$$

$$\frac{x}{a} + \frac{z}{c} = v\left(1 - \frac{y}{b}\right), \quad v\left(\frac{x}{a} - \frac{z}{c}\right) = 1 + \frac{y}{b}, \tag{3.422b}$$

where u and v are arbitrary quantities.

b) Hyperbolic Paraboloid (Fig. 3.197):

$$z = \frac{x^2}{a^2} - \frac{y^2}{b^2}. \tag{3.423}$$

The hyperbolic paraboloid also has two families of rectilinear generators with equations

$$\frac{x}{a} + \frac{y}{b} = u, \ u\left(\frac{x}{a} - \frac{y}{b}\right) = z; \tag{3.424a}$$ $$\frac{x}{a} - \frac{y}{b} = v, \ v\left(\frac{x}{a} + \frac{y}{b}\right) = z. \tag{3.424b}$$

The quantities u and v are again arbitrary values. In both cases, there are two straight lines passing through every point of the surface, one from each family. Only one family of straight lines is denoted in **Fig. 3.196** and **Fig. 3.197**.

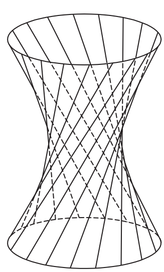


Figure 3.196

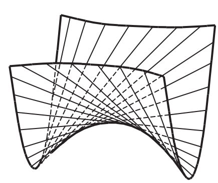


Figure 3.197

7. Cylinder

a) Elliptic Cylinder (Fig. 3.198): $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1.$ (3.425)

b) Hyperbolic Cylinder (Fig. 3.199): $\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1.$ (3.426)

c) Parabolic Cylinder (Fig. 3.200): $y^2 = 2px.$ (3.427)

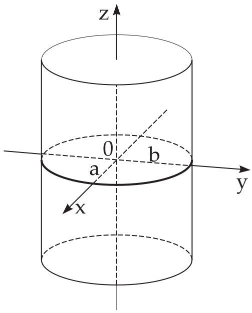


Figure 3.198

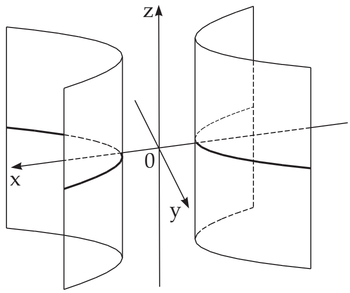


Figure 3.199

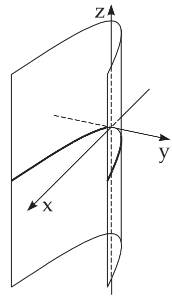


Figure 3.200

3.5.3.14 Surfaces of Second Order or Quadratic Surfaces, General Theory

1. General Equation of a Surface of Second Order

$$a_{11}x^2 + a_{22}y^2 + a_{33}z^2 + 2a_{12}xy + 2a_{23}yz + 2a_{31}zx + 2a_{14}x + 2a_{24}y + 2a_{34}z + a_{44} = 0. \quad (3.428)$$

2. Telling the Type of Second-Order Surface from its Equation

The type of a second-order surface can be determined from its equation by the signs of its invariants Δ, δ, S, and T from **Tables 3.24** and **3.25**. Here one can find the names with the normal form of the equation of the surfaces, and every equation can be transformed into a normal form. From the equation of the so-called imaginary surfaces it is impossible to determine the coordinates of any real point, except the vertex of the imaginary cone, and the intersection line of two imaginary planes.

Table 3.24 Type of surfaces of second order with $\delta \neq 0$ (central surfaces)

	$\boldsymbol{S \cdot \delta > 0, \quad T > 0}$	$\boldsymbol{S \cdot \delta}$ **and** $\boldsymbol{T}$ **not both** $\boldsymbol{> 0}$
$\boldsymbol{\Delta < 0}$	Ellipsoid $\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1$	Hyperboloid of two sheets $\frac{x^2}{a^2} + \frac{y^2}{b^2} - \frac{z^2}{c^2} = -1$
$\boldsymbol{\Delta > 0}$	Imaginary ellipsoid $\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = -1$	Hyperboloid of one sheet $\frac{x^2}{a^2} + \frac{y^2}{b^2} - \frac{z^2}{c^2} = 1$
$\boldsymbol{\Delta = 0}$	Imaginary cone (with real vertex) $\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 0$	Cone $\frac{x^2}{a^2} + \frac{y^2}{b^2} - \frac{z^2}{c^2} = 0$

3. Invariants of a Surface of Second Order

Substituting $a_{ik} = a_{ki}$, then holds:

$$\Delta = \begin{vmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{vmatrix}, \quad (3.429a) \qquad \delta = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}, \quad (3.429b)$$

$$S = a_{11} + a_{22} + a_{33}, \quad (3.429c) \qquad T = a_{22}a_{33} + a_{33}a_{11} + a_{11}a_{22} - a_{23}^2 - a_{31}^2 - a_{12}^2. \quad (3.429d)$$

During translation or rotation of the coordinate system these invariants do not change.

Table 3.25 Type of surfaces of second order with $\delta = 0$ (paraboloid, cylinder and two planes)

	$\boldsymbol{\Delta < 0}$ (here $\boldsymbol{T > 0}$)[1]	$\boldsymbol{\Delta > 0}$ (here $\boldsymbol{T < 0}$)
$\boldsymbol{\Delta \neq 0}$	Elliptic paraboloid $\frac{x^2}{a^2} + \frac{y^2}{b^2} = \pm z$	Hyperbolic paraboloid $\frac{x^2}{a^2} - \frac{y^2}{b^2} = \pm z$
$\boldsymbol{\Delta = 0}$	Cylindrical surface with a second-order curve as a directrix whose type defines different cylinders: For $T > 0$ imaginary elliptic, for $T < 0$ hyperbolic, and for $T = 0$ parabolic cylinder, if the surface does not split into two real, imaginary, or coincident planes. The condition for splitting is: $\begin{vmatrix} a_{11} & a_{12} & a_{14} \\ a_{21} & a_{22} & a_{24} \\ a_{41} & a_{42} & a_{44} \end{vmatrix} + \begin{vmatrix} a_{11} & a_{13} & a_{14} \\ a_{31} & a_{33} & a_{34} \\ a_{41} & a_{43} & a_{44} \end{vmatrix} + \begin{vmatrix} a_{22} & a_{23} & a_{24} \\ a_{32} & a_{33} & a_{34} \\ a_{42} & a_{43} & a_{44} \end{vmatrix} = 0$	

[1] For the quantities Δ and T see p. 228.

3.5.4 Geometric Transformations and Coordinate Transformations

1. Transformations

Transformations describe the changes of the position or of the form of objects in the plane or in the space. There are two types of consideration which are in close relationship with each other [3.22]. In the first case points or objects are transformed in a fixed coordinate system. It is called *geometric transformation* (3.5.4.1, p. 229, 3.5.4.5,**1.**, p. 234).

In the second case the object remains unchanged while the coordinate system is transformed with respect to the object. With this *coordinate transformation* (3.5.4.3, p. 231, 3.5.4.5,**2.**, p. 234) not the object, but its coordinate representations are modified. At problem solving one of them can be more reasonable than the other.

2. Application Fields

- Building parts which are described in their own *object coordinate-system.*
- Description of motions of parts connected to each other (e.g. robots).
- Reproduction of two-dimensional projections of three-dimensional objects.
- Description of motions and deformations in computer graphics and computer animation.

3.5.4.1 Geometric 2D Transformations

1. Translations

Translation of a point P, given by Cartesian coordinates x_P, y_P, by t_x in the direction of the positive x-axis, and by t_y in the direction of the positive y-axis **(Fig.3.201)** results the new coordinates of the transformed point $P'(x'_P, y'_P)$:

$$x'_P = x_P + t_x\,, \qquad y'_P = y_P + t_y\,. \tag{3.430}$$

If the coordinates are described as column vectors, then the formulas for the new coordinates of the transformation and the reversal are

$$\begin{pmatrix} x'_P \\ y'_P \end{pmatrix} = \begin{pmatrix} x_P \\ y_P \end{pmatrix} + \begin{pmatrix} t_x \\ t_y \end{pmatrix}, \quad (3.431a) \qquad \begin{pmatrix} x_P \\ y_P \end{pmatrix} = \begin{pmatrix} x'_P \\ y'_P \end{pmatrix} - \begin{pmatrix} t_x \\ t_y \end{pmatrix}. \quad (3.431b)$$

2. Rotation Around the Origin

At a rotation the object is rotated around the origin by an angle α. If $\alpha > 0$, then the rotation is counterclockwise **(Fig.3.202)**. The mapping of the coordinates of a point $P(x_P, y_P)$ is described by

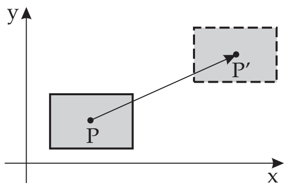


Figure 3.201

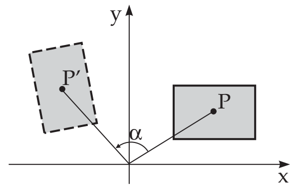


Figure 3.202

the relations

$$x'_P = x_P \cdot \cos\alpha - y_P \cdot \sin\alpha\,, \qquad y'_P = x_P \cdot \sin\alpha + y_P \cdot \cos\alpha. \tag{3.432}$$

The matrix form of (3.432) is

$$\begin{pmatrix} x'_P \\ y'_P \end{pmatrix} = \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix} \begin{pmatrix} x_P \\ y_P \end{pmatrix}. \tag{3.433a}$$

The inverse of this transformation corresponds to a rotation by the angle $-\alpha$:

$$\begin{pmatrix} x_P \\ y_P \end{pmatrix} = \begin{pmatrix} \cos(-\alpha) & -\sin(-\alpha) \\ \sin(-\alpha) & \cos(-\alpha) \end{pmatrix} \begin{pmatrix} x'_P \\ y'_P \end{pmatrix} = \begin{pmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{pmatrix} \begin{pmatrix} x'_P \\ y'_P \end{pmatrix}. \tag{3.433b}$$

3. Scaling with respect to the Origin

At scaling transformations the coordinates are multiplied by s_x and s_y respectively **(Fig.3.203)**. The transformation of a point $P(x_P, y_P)$ is given by

$$x'_P = s_x \cdot x_P\,, \qquad y'_P = s_y \cdot y_P\,. \tag{3.434}$$

The matrix forms of this transformation and of its inverse are

$$\begin{pmatrix} x'_P \\ y'_P \end{pmatrix} = \begin{pmatrix} s_x & 0 \\ 0 & s_y \end{pmatrix} \begin{pmatrix} x_P \\ y_P \end{pmatrix}, \tag{3.435a}$$

$$\begin{pmatrix} x_P \\ y_P \end{pmatrix} = \begin{pmatrix} 1/s_x & 0 \\ 0 & 1/s_y \end{pmatrix} \begin{pmatrix} x'_P \\ y'_P \end{pmatrix}. \tag{3.435b}$$

Scaling results in a change of the size of the transformed object. A positive multiplier $s_x < 1$ results in contraction of the object in x-direction. Consequently a factor $s_x > 1$ results in an expansion. A negative factor $s_x < 0$ results in a reflection with respect to the y-axis. The corresponding statements are valid for s_y.

Scaling transformations in special cases:
- Reflection with respect to the x-axis: $s_x = 1$, $s_y = -1$.
- Reflection with respect to the y-axis: $s_x = -1$, $s_y = 1$.
- Reflection with respect to the origin: $s_x = s_y = -1$.

4. Shearing

At shear transformation the value of each coordinate is changed proportionally to the other. The formulas of this transformation are:

$$x'_P = x_P + a_x \cdot y_P\,, \qquad y'_P = y_P + a_y \cdot x_P\,. \tag{3.436}$$

The matrix form of this transformation (with notation $m = 1 - a_x a_y$) is seen in the following:

$$\begin{pmatrix} x'_P \\ y'_P \end{pmatrix} = \begin{pmatrix} 1 & a_x \\ a_y & 1 \end{pmatrix} \begin{pmatrix} x_P \\ y_P \end{pmatrix}, \tag{3.437a}$$

$$\begin{pmatrix} x_P \\ y_P \end{pmatrix} = \begin{pmatrix} 1/m & -a_x/m \\ -a_y/m & 1/m \end{pmatrix} \begin{pmatrix} x'_P \\ y'_P \end{pmatrix}. \tag{3.437b}$$

Fig.3.204 shows an example for shearing.

5. Properties of Transformations

The transformations introduced above are *affine transformations*, i.e. the (x', y') coordinates of a transformed point P' can be expressed by a system of linear equations of the original coordinates (x, y) of P.

Remark: These transformations keep collinearity and parallelism, i.e. lines are transformed into lines,

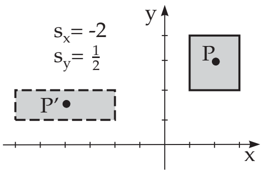


Figure 3.203

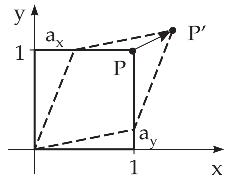


Figure 3.204

and the images of parallel lines are parallel lines. Furthermore translation, rotation and reflection are distance- and angle-preserving mappings.

3.5.4.2 Homogeneous Coordinates, Matrix Representation

While the changes of coordinates in the case of rotation, scaling and shearing can be described by multiplication by a 2×2 type matrix (3.433a),(3.435a) and (3.437a), translation does not have this type of representation. To be able to handle all of these transformations in the same way, *homogeneous coordinates* are introduced. Every point in the plane gets an additional coordinate $w \neq 0$. Point $P(x, y)$ will have coordinates (x^h, y^h, w), where

$$x = \frac{x^h}{w}, \qquad y = \frac{y^h}{w}. \tag{3.438}$$

In the followings w is fixed as $w = 1$. So point $P(x, y)$ has coordinates $(x, y, 1)$. Now the basic transformations can be given by a 3×3 type matrix in the following form:

$$\begin{pmatrix} x'_P \\ y'_P \\ 1 \end{pmatrix} = \begin{pmatrix} m_{11} & m_{12} & m_{13} \\ m_{21} & m_{22} & m_{23} \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_P \\ y_P \\ 1 \end{pmatrix}, \quad \text{i.e., } \underline{x}_P' = \mathbf{M}\underline{x}_P \text{ holds}. \tag{3.439}$$

The matrices for translation, rotation, scaling and shearing are:

$$\mathbf{T}(t_x, t_y) = \begin{pmatrix} 1 & 0 & t_x \\ 0 & 1 & t_y \\ 0 & 0 & 1 \end{pmatrix}, \tag{3.440}$$

Translation matrix

$$\mathbf{R}(\alpha) = \begin{pmatrix} \cos\alpha & -\sin\alpha & 0 \\ \sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{pmatrix}, \tag{3.441}$$

Rotation matrix

$$\mathbf{S}(s_x, s_y) = \begin{pmatrix} s_x & 0 & 0 \\ 0 & s_y & 0 \\ 0 & 0 & 1 \end{pmatrix}, \tag{3.442}$$

Scaling matrix

$$\mathbf{V}(a_x, a_y) = \begin{pmatrix} 1 & a_x & 0 \\ a_y & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{3.443}$$

Shearing matrix

3.5.4.3 Coordinate Transformation

With geometric transformations an object is transformed with respect to a fixed coordinate system. Whereas *coordinate transformation* gives the relation between the coordinate representations of a fixed object with respect to two different coordinate systems.

The relation between the two types of transformations is represented in **Fig.3.205**. If the coordinate system is shifted by vector $\vec{\mathbf{t}}$, the coordinates of point $P(x_P, y_P)$ become $x'_P = x_P - t_x$, $y'_P = y_P - t_y$. The translation of the coordinate system by the vector $\vec{\mathbf{t}}$ results the same outcome as the translation of point P by $-\vec{\mathbf{t}}$.

The same consequences are valid for rotation and scaling. So, the transformation of the coordinate system is equivalent to the reverse transformation of the object.

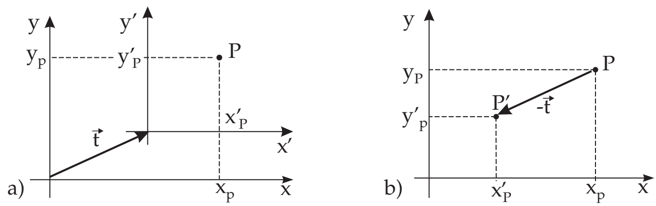


Figure 3.205

The coordinate transformations resulted by a translation, rotation or scaling can be given by 3×3 transformation matrices $\overline{\mathbf{T}}$, $\overline{\mathbf{R}}$ and $\overline{\mathbf{S}}$. Considering the matrices of geometric transformations (3.440)–(3.442) follows:

$$\overline{\mathbf{T}}(t_x, t_y) = \mathbf{T}(-t_x, -t_y) = \mathbf{T}^{-1}(t_x, t_y) \quad (3.444) \qquad \overline{\mathbf{R}}(\alpha) = \mathbf{R}(-\alpha) = \mathbf{R}^{-1}(\alpha) \quad (3.445)$$

$$\overline{\mathbf{S}}(s_x, s_y) = \mathbf{S}(1/s_x, 1/s_y) = \mathbf{S}^{-1}(s_x, s_y) \quad (3.446)$$

In this way all basic transformations can be described by a 3×3 transformation matrix $\overline{\mathbf{M}}$:

$$\begin{pmatrix} x'_P \\ y'_P \\ 1 \end{pmatrix} = \begin{pmatrix} \overline{m}_{11} & \overline{m}_{12} & \overline{m}_{13} \\ \overline{m}_{21} & \overline{m}_{22} & \overline{m}_{23} \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_P \\ y_P \\ 1 \end{pmatrix}, \quad \text{i.e. } \underline{x}'_{\mathrm{P}} = \overline{\mathbf{M}}\underline{x}_{\mathrm{P}}\,. \quad (3.447)$$

3.5.4.4 Composition of Transformations

Complex geometric transformations can be realized by combining different basic transformations. Let a sequence of transformations be given by matrices $\mathbf{M}_1$, $\mathbf{M}_2$,...,$\mathbf{M}_n$. The consecutive execution of these transformations converts the point $P(x, y)$ in n steps into P'. The transformation matrix $\mathbf{M}$ resulted by this sequence of mappings is the product of these matrices:

$$\mathbf{M} = \mathbf{M}_n \cdot \mathbf{M}_{n-1} \cdot \ldots \cdot \mathbf{M}_2 \cdot \mathbf{M}_1. \quad (3.448)$$

Similarly, for the reverse transformation

$$\mathbf{M}^{-1} = \mathbf{M}_1^{-1} \cdot \mathbf{M}_2^{-1} \cdot \ldots \cdot \mathbf{M}_{n-1}^{-1} \cdot \mathbf{M}_n^{-1}. \quad (3.449)$$

So, instead of performing n times basic transformations in a sequence on a point, it is possible to give the matrix of the composite transformation, and to apply it directly.
Every affine transformation can be given as a chain (composition) of translation, rotation and scaling. Even shearing can be given as a rotation $\mathbf{R}(\alpha)$, a scaling $\mathbf{S}(s_x, s_y)$ and a further rotation $\mathbf{R}(\beta)$ applied after each other. The parameters $\alpha, \beta, s_x,\ s_y$ can be determined so, that $\mathbf{V}(a, b) = \mathbf{R}(\beta)\cdot\mathbf{S}(s_x, s_y)\cdot\mathbf{R}(\alpha)$ holds.

■ Calculation of the transformation matrix of a rotation around an arbitrary point $Q(x_q, y_q)$ by an angle α: The composite transformation is the result of composition of the following basic transformations:
1. Shifting of Q into the origin: $\mathbf{M}_1 = \mathbf{T}(-x_q, -y_q)$.
2. Rotation around the origin: $\mathbf{M}_2 = \mathbf{R}(\alpha)$.
3. Translation back the origin into Q: $\mathbf{M}_3 = \mathbf{M}_1^{-1} = \mathbf{T}(x_q, y_q)$.
The sequence of the single steps of these transformations see **Fig.3.206**. Point P is transformed via P_1 and P_2 into P'.

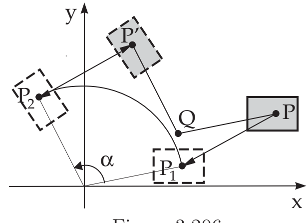


Figure 3.206

$$\mathbf{M} = \mathbf{M}_3 \cdot \mathbf{M}_2 \cdot \mathbf{M}_1 = \mathbf{T}(x_q, y_q) \cdot \mathbf{R}(\alpha) \cdot \mathbf{T}(-x_q, -y_q)$$

$$= \begin{pmatrix} 1 & 0 & x_q \\ 0 & 1 & y_q \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \cos\alpha & -\sin\alpha & 0 \\ \sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & -x_q \\ 0 & 1 & -y_q \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} \cos\alpha & -\sin\alpha & x_q(1-\cos\alpha) + y_q\sin\alpha \\ \sin\alpha & \cos\alpha & y_q(1-\cos\alpha) - x_q\sin\alpha \\ 0 & 0 & 1 \end{pmatrix}.$$

■ Reflection with respect to a line given by the equation $y = mx + n$:
1. Translation of the line to pass through the origin: $\mathbf{M}_1 = \mathbf{T}(0, -n)$.
2. Rotation of the line clockwise until it coincides with the x-axis: $\mathbf{M}_2 = \mathbf{R}(-\alpha)$, with $\tan\alpha = m$.
3. Reflection with respect to the x-axis: $\mathbf{M}_3 = \mathbf{S}(1, -1)$.
4. Rotation back by α: $\mathbf{M}_4 = \mathbf{M}_2^{-1} = \mathbf{R}(\alpha)$.
5. Translation of the line back into the original position: $\mathbf{M}_5 = \mathbf{M}_1^{-1} = \mathbf{T}(0, n)$.

$$\mathbf{M} = \mathbf{T}(0,n)\cdot\mathbf{R}(\alpha)\cdot\mathbf{S}(1,-1)\cdot\mathbf{R}(-\alpha)\cdot\mathbf{T}(0,-n)$$

$$= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & n \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \cos\alpha & -\sin\alpha & 0 \\ \sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \cos\alpha & \sin\alpha & 0 \\ -\sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & -n \\ 0 & 0 & 1 \end{pmatrix}.$$

Use of the well known trigonometric relations $\sin\alpha = m/\sqrt{m^2+1}$ and $\cos\alpha = 1/\sqrt{m^2+1}$ results the transformation matrix

$$\mathbf{M} = \begin{pmatrix} \dfrac{1-m^2}{m^2+1} & \dfrac{2m}{m^2+1} & \dfrac{-2mn}{m^2+1} \\ \dfrac{2m}{m^2+1} & \dfrac{m^2-1}{m^2+1} & \dfrac{2n}{m^2+1} \\ 0 & 0 & 1 \end{pmatrix}.$$

■ Complete centered transformation of a rectangle shape cut with side lengths a and b into a similar one in a window with width c and height d **(Fig.3.207)**. Sequence of the basic transformations:
1. Shifting of $P(x_P, y_P)$ into the origin: $\mathbf{M}_1 = \mathbf{T}(-x_P, -y_P)$.
2. Clockwise rotation by an angle α: $\mathbf{M}_2 = \mathbf{R}(-\alpha)$.
3. Scaling by factor $s = s_x = s_y = \min(c/a, d/b)$: $\mathbf{M}_3 = \mathbf{S}(s, s)$
4. Shifting of the origin into the center of the window: $\mathbf{M}_4 = \mathbf{T}(c/2, d/2)$.

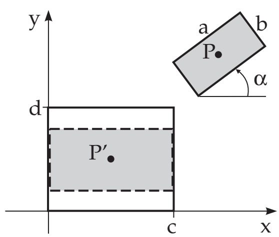


Figure 3.207

$$\mathbf{M} = \mathbf{T}(c/2, d/2)\cdot\mathbf{S}(s,s)\cdot\mathbf{R}(-\alpha)\cdot\mathbf{T}(-x_P, -y_P)$$

$$= \begin{pmatrix} 1 & 0 & c/2 \\ 0 & 1 & d/2 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} s & 0 & 0 \\ 0 & s & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \cos\alpha & \sin\alpha & 0 \\ -\sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & -x_P \\ 0 & 1 & -y_P \\ 0 & 0 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} s\cos\alpha & s\sin\alpha & c/2 - s(x_P\cos\alpha + y_P\sin\alpha) \\ -s\sin\alpha & s\cos\alpha & d/2 - s(y_P\cos\alpha - x_P\sin\alpha) \\ 0 & 0 & 1 \end{pmatrix}.$$

3.5.4.5 3D–Transformations

The mathematical description of geometric transformations and coordinate transformations in the three-dimensional space is based on the same ideas which have been discussed for the two-dimensional case in sections 3.5.4.1–3.5.4.4. Affine transformations of the three-dimensional space are the compositions of basic transformations as: translation, rotation around one of the coordinate axes and scaling with respect to the origin. Using homogeneous coordinates these transformations can be given by 4×4 transformation matrices. As in the two-dimensional case, the composed transformations can be realized by matrix multiplications.

1. Geometric Transformations

The transformation of a point $P(x_P, y_P, z_P)$ is performed according to the following rule:

$$\begin{pmatrix} x'_P \\ y'_P \\ z'_P \\ 1 \end{pmatrix} = \begin{pmatrix} m_{11} & m_{12} & m_{13} & m_{14} \\ m_{21} & m_{22} & m_{23} & m_{24} \\ m_{31} & m_{32} & m_{33} & m_{34} \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_P \\ y_P \\ z_P \\ 1 \end{pmatrix}, \quad \text{i.e. } \underline{x}_{\text{P}}' = \mathbf{M}\underline{x}_{\text{P}}\,. \tag{3.450}$$

The transformation matrices of the basic transformations are:

$$\mathbf{T}(t_x, t_y, t_z) = \begin{pmatrix} 1 & 0 & 0 & t_x \\ 0 & 1 & 0 & t_y \\ 0 & 0 & 1 & t_z \\ 0 & 0 & 0 & 1 \end{pmatrix}, \tag{3.451}$$

Translation

$$\mathbf{S}(s_x, s_y, s_z) = \begin{pmatrix} s_x & 0 & 0 & 0 \\ 0 & s_y & 0 & 0 \\ 0 & 0 & s_z & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \tag{3.452}$$

Scaling with respect to the origin

$$\mathbf{R}_x(\alpha) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos\alpha & -\sin\alpha & 0 \\ 0 & \sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \tag{3.453}$$

Rotation around the x-axis

$$\mathbf{R}_y(\alpha) = \begin{pmatrix} \cos\alpha & 0 & \sin\alpha & 0 \\ 0 & 1 & 0 & 0 \\ -\sin\alpha & 0 & \cos\alpha & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \tag{3.454}$$

Rotation around the y-axis

$$\mathbf{R}_z(\alpha) = \begin{pmatrix} \cos\alpha & -\sin\alpha & 0 & 0 \\ \sin\alpha & \cos\alpha & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \tag{3.455}$$

Rotation around the z-axis

$$\mathbf{V}_{xy}(a_x, a_y) = \begin{pmatrix} 1 & 0 & a_x & 0 \\ 0 & 1 & a_y & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{3.456}$$

Shearing parallel to the x, y-plane

For a positive α the rotation is counterclockwise looking from the positive half of the coordinate axis to the origin. For reverse transformations the following relations are valid:

$$\mathbf{T}^{-1}(t_x, t_y, t_z) = \mathbf{T}(-t_x, -t_y, -t_z), \quad \mathbf{S}^{-1}(s_x, s_y, s_z) = \mathbf{S}(1/s_x, 1/s_y, 1/s_z)\,, \tag{3.457}$$

$$\mathbf{R}_x^{-1}(\alpha) = \mathbf{R}_x(-\alpha), \quad \mathbf{R}_y^{-1}(\alpha) = \mathbf{R}_y(-\alpha), \quad \mathbf{R}_z^{-1}(\alpha) = \mathbf{R}_z(-\alpha)\,. \tag{3.458}$$

2. Coordinate Transformations

Analogously to the two-dimensional case, the transformations of the coordinate systems have the same effects to the coordinate representations of a point as the reverse geometric transformation (see 3.5.4.3, p. 231). Hence, the matrices of the transformations are:

$$\overline{\mathbf{T}}(t_x, t_y, t_z) = \mathbf{T}(-t_x, -t_y, -t_z) = \mathbf{T}^{-1}(t_x, t_y, t_z)\,, \tag{3.459}$$

$$\overline{\mathbf{R}}_x(\alpha_x) = \mathbf{R}_x(-\alpha_x) = \mathbf{R}_x^{-1}(\alpha_x)\,, \tag{3.460}$$

$$\overline{\mathbf{R}}_y(\alpha_y) = \mathbf{R}_y(-\alpha_y) = \mathbf{R}_y^{-1}(\alpha_y)\,, \tag{3.461}$$

$$\overline{\mathbf{R}}_z(\alpha_z) = \mathbf{R}_z(-\alpha_z) = \mathbf{R}_z^{-1}(\alpha_z)\,, \tag{3.462}$$

$$\overline{\mathbf{S}}(s_x, s_y, s_z) = \mathbf{S}(1/s_x, 1/s_y, 1/s_z) = \mathbf{S}^{-1}(s_x, s_y, s_z)\,. \tag{3.463}$$

In practical applications it occurs quite often, that a particular transformation is replaced from a right-handed Cartesian coordinate system into another Cartesian coordinate system. The original one is often called as *world coordinate system*, the other one is called the *local* or *object coordinate system* . If the origin $U(x_u, y_u, z_u)$ and the unit vectors $\vec{\mathbf{e}}_i = \{l_1, m_1, n_1\}$, $\vec{\mathbf{e}}_j = \{l_2, m_2, n_2\}$ and $\vec{\mathbf{e}}_k = \{l_3, m_3, n_3\}$ of the local coordinate system are given in the world coordinate system, then the transformation from the world system into the local system and its reverse are given by the matrices

$$\overline{\mathbf{M}} = \begin{pmatrix} l_1 & m_1 & n_1 & -l_1x_u - m_1y_u - n_1z_u \\ l_2 & m_2 & n_2 & -l_2x_u - m_2y_u - n_2z_u \\ l_3 & m_3 & n_3 & -l_3x_u - m_3y_u - n_3z_u \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad (3.464) \qquad \overline{\mathbf{M}}^{-1} = \begin{pmatrix} l_1 & l_2 & l_3 & x_u \\ m_1 & m_2 & m_3 & y_u \\ n_1 & n_2 & n_3 & z_u \\ 0 & 0 & 0 & 1 \end{pmatrix}. \quad (3.465)$$

If a point P has the coordinates (x_P, y_P, z_P) in the world coordinate system and it has the coordinates (x'_P, y'_P, z'_P) in the local system, then the following equalities are valid:

$$\underline{x}_P{}' = \overline{\mathbf{M}}\underline{x}_P\,, \quad (3.466) \qquad \underline{x}_P = \overline{\mathbf{M}}^{-1}\underline{x}_P{}'\,. \quad (3.467)$$

If $\overline{\mathbf{M}}_1$ and $\overline{\mathbf{M}}_2$ denote the matrices of transformations from the world coordinate system into two local coordinate systems, then the transformation between the two local systems are given by the matrices:

$$\overline{\mathbf{M}} = \overline{\mathbf{M}}_1 \cdot \overline{\mathbf{M}}_2^{-1} \quad \text{and} \quad \overline{\mathbf{M}}^{-1} = \overline{\mathbf{M}}_2 \cdot \overline{\mathbf{M}}_1^{-1}. \quad (3.468)$$

■ Determination of the matrix of a rotation by an angle θ around the line passing through the points $P(x_P, y_P, z_P)$ and $Q(x_q, y_q, z_q)$ with direction vector $\vec{\mathbf{v}} = \{v_x, v_y, v_z\} = \{x_P - x_q, y_P - y_q, z_P - z_q\}$. It is easy to choose P and Q in a distance of one unit, so $\vec{\mathbf{v}}$ is a unit vector. First the line is transformed into the z-axis of the coordinate system. Then the rotation follows, by an angle θ around the z-axis. Finally it follows the transformation of the line back into the original one. It is represented in **Fig.3.208** how the transformation of the spatial line into the z-axis is performed. It consists of the following steps:

1. Q is translated into the origin: $\quad \mathbf{M}_1 = \mathbf{T}(-x_q, -y_q, -z_q)$.

2. Rotation around the z-axis so that the rotation-axis is mapped into the y,z-plane: $\mathbf{M}_2 = \mathbf{R}_z(\alpha_z)$ with $\cos\alpha_z = v_y\Big/\sqrt{v_x^2 + v_y^2}$ and $\sin\alpha_z = v_x\Big/\sqrt{v_x^2 + v_y^2}$.

Point P_2 has the coordinates $(0, \sqrt{v_x^2 + v_z^2}, v_z)$.

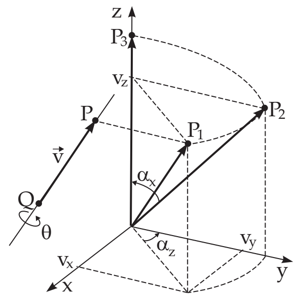


Figure 3.208

3. Rotation around the x-axis by the angle α_x until the image of the direction vector $\vec{\mathbf{v}}$ is on the z-axis: $\mathbf{M}_3 = \mathbf{R}_x(\alpha_x)$, where $\cos\alpha_x = v_z/|\vec{\mathbf{v}}|$ and $\sin\alpha_x = \sqrt{v_x^2 + v_y^2}\Big/|\vec{\mathbf{v}}|$. The point P_3 has the coordinates $(0, 0, |\vec{\mathbf{v}}|)$.

The matrix of transforming the direction vector into the z-axis is with $m = \sqrt{v_x^2 + v_y^2}$ and $|\vec{\mathbf{v}}| = 1$:

$$\mathbf{M}_A = \mathbf{R}_x(\alpha_x) \cdot \mathbf{R}_z(\alpha_z) \cdot \mathbf{T}(-x_q, -y_q, -z_q)$$

$$= \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & v_z & -m & 0 \\ 0 & m & v_z & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \frac{v_y}{m} & \frac{-v_x}{m} & 0 & 0 \\ \frac{v_x}{m} & \frac{v_y}{m} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & -x_q \\ 0 & 1 & 0 & -y_q \\ 0 & 0 & 1 & -z_q \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

$$\mathbf{M}_A = \begin{pmatrix} \frac{v_y}{m} & \frac{-v_x}{m} & 0 & \frac{v_xy_q - v_yx_q}{m} \\ \frac{v_xv_z}{m} & \frac{v_yv_z}{m} & -m & mz_q - \frac{v_xv_zx_q + v_yv_zy_q}{m} \\ v_x & v_y & v_z & -v_xx_q - v_yy_q - v_zz_q \\ 0 & 0 & 0 & 1 \end{pmatrix}, \qquad \mathbf{M}_A^{-1} = \begin{pmatrix} \frac{v_y}{m} & \frac{v_xv_z}{m} & v_x & x_q \\ \frac{-v_x}{m} & \frac{v_yv_z}{m} & v_y & y_q \\ 0 & -m & v_z & z_q \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Comparing $\mathbf{M}_A$ and $\mathbf{M}_A^{-1}$ with matrices (3.465) and (3.464) shows, that the transformation of the spatial line into the z-axis is identical to a coordinate transformation from the world coordinate system into a local coordinate system where the origin is Q and the z-axis-direction is $\vec{\mathbf{v}}$. In the local system the rotation is made around the z-axis. The matrix of transformation of the complete rotation is given by matrix (3.455)

$$\mathbf{M} = \mathbf{M}_A^{-1} \cdot \mathbf{R}_z(\theta) \cdot \mathbf{M}_A$$

$$= \begin{pmatrix} \frac{v_y}{m} & \frac{v_x v_z}{m} & v_x & x_q \\ \frac{-v_x}{m} & \frac{v_y v_z}{m} & v_y & y_q \\ 0 & -m & v_z & z_q \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \cos\theta & -\sin\theta & 0 & 0 \\ \sin\theta & \cos\theta & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \frac{v_y}{m} & \frac{-v_x}{m} & 0 & \frac{v_x y_q - v_y x_q}{m} \\ \frac{v_x v_z}{m} & \frac{v_y v_z}{m} & -m & m z_q - \frac{v_x v_z x_q + v_y v_z y_q}{m} \\ v_x & v_y & v_z & -v_x x_q - v_y y_q - v_z z_q \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

In section 4.4, p. 289 an alternative method is given to describe rotation transformations by using the properties of *quaternions*.

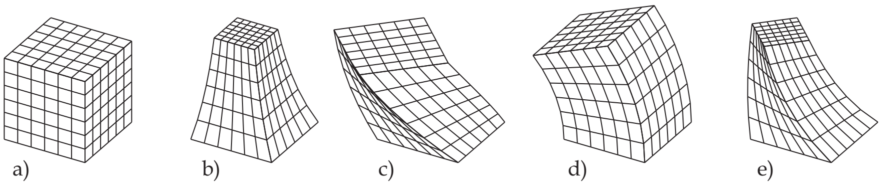


Figure 3.209

3.5.4.6 Deformation Transformations

The affine transformations discussed in section 3.5.4.5, p. 233 change the position of objects and result in stretching or compressing in given directions. If the elements of the transformation matrix $\mathbf{M} = (m_{ij})$ are not constants, as they have been until now, but they are functions of position, then this will lead to a generalized class of structure-changing transformations. The elements of the matrix now have the form:

$$m_{ij} = m_{ij}(x, y, z). \tag{3.469}$$

1. Contraction

This transformation is a generalization of scaling. At a contraction in the direction of the z-axis the scaling parameters s_x, s_y are functions of z. The matrix of transformation is:

$$\mathbf{S}(s_x(z), s_y(z), 1) = \begin{pmatrix} s_x(z) & 0 & 0 & 0 \\ 0 & s_y(z) & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{3.470}$$

The functions $s_x(z)$ and $s_y(z)$ define the contraction-profile. If $s_x(z) > 1$, then the transformed object is stretched in the direction of the x-axis, if $s_x(z) < 1$, then the object is compressed. **Fig.3.209b** shows the result of the transformation of the unit-cube in **Fig.3.209a** by functions $s_x(z) = s_y(z) = 1/(1-z^2)$.

2. Torsion around the z-axis

This transformation is the generalization of the rotation around the z-axis. The angle of rotation is changing along the z-axis. The matrix of transformation is:

$$\mathbf{R}_z(\alpha(z)) = \begin{pmatrix} \cos\alpha(z) & -\sin\alpha(z) & 0 & 0 \\ \sin\alpha(z) & \cos\alpha(z) & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{3.471}$$

The function $\alpha(z)$ defines the angle of rotation along the z-axis. **Fig.3.209c** shows the rotation of the unit-cube with $\alpha(z) = \frac{\pi}{4}z$.

3. Bending Around the x-axis

At bending the angle of rotation changes in a direction perpendicular to the rotation-axis. The matrix of transformation has the following form:

$$\mathbf{R}_x(\alpha(z)) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos\alpha(z) & -\sin\alpha(z) & 0 \\ 0 & \sin\alpha(z) & \cos\alpha(z) & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{3.472}$$

Bending of the unit-cube by $\alpha(z) = \frac{\pi}{8}z$ is shown in **Fig.3.209d**.

These deformations can be applied in a sequence like affine transformations. The object in **Fig.3.209e** is the result of a contraction then a torsion of the unit-cube.

3.5.5 Planar Projections

There are several methods to visualize three-dimensional objects in a two-dimensional media [3.22]. Among them planar projections are very important. A *planar projection* is a mapping, where points of a three-dimensional space are assigned to points of a plane. An *image point* is given as the intersection point of this plane and a ray connecting an observer with the space-point. The plane is called a *projection plane* or a *picture plane*, the ray is called *projecting ray*, and its direction is the *direction of the projection* .

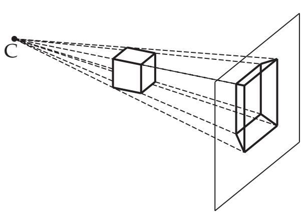


Figure 3.210

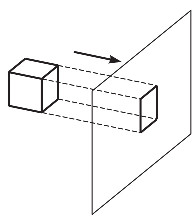


Figure 3.211

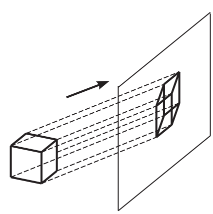


Figure 3.212

3.5.5.1 Classification of the projections

1. Central Projection

The *central projection* which is called also *perspective projection* the projecting rays start from a common center point **(Fig.3.210)**. Objects being farther from the center of perspectivity C are represented smaller than the ones being closer to the center. Parallel lines, which are not parallel to the projection plane, become not parallel and they intersect in a so called *vanishing point*. Perspective projections give a realistic impression of the object for the viewer. But relations of lengths and angles are lost at this mapping.

2. Parallel Projection

At the *parallel projection* the projecting rays are parallel to each other **(Fig.3.211)**. Line segments, which are not parallel to the projection plane are shortened, and angles are usually distorted.

1. Orthogonal Parallel Projections A parallel projection is an *orthogonal projection* if the direction of the projecting ray is perpendicular to the picture plane. If furthermore the picture plane is perpendicular to one of the coordinate-axes, then it is an *orthographic projection* or principal projection, which is well known from industrial designs.

If the direction of projection is not perpendicular to any coordinate-axis, then the orthogonal projection is called an *axonometric projection*.

2. Oblique Parallel Projections A parallel projection is an *oblique projection* if the direction of the projecting ray is not parallel to the normal vector of the picture plane **(Fig.3.212)**. Special cases of oblique projections are *Cavalier projection* and *cabinet projection* .

Sometimes parallel projections preserve the ratio of sizes, but they seem less realistic than perspective representations.

3.5.5.2 Local or Projection Coordinate System

Defining the direction of the projection plane and the coordinates of the result of a projection in the world coordinate system is not reasonable. It seems to be useful to apply a picture coordinate system whose x, y-plane is identical to the projection plane. This picture system represents the projection from the viewpoint of an observer who is looking perpendicularly to the projection plane.

Let the projection plane be given by a reference point $R(x_r, y_r, z_r)$ and a unit normalvector $\vec{\mathbf{n}} = \{n_x, n_y, n_z\}$. Then the picture coordinate system is defined in the following way. The origin is placed into $R(x_r, y_r, z_r)$. Among the coordinate unit vectors $\vec{\mathbf{e}}'_i$, $\vec{\mathbf{e}}'_j$ and $\vec{\mathbf{e}}'_k$ first $\vec{\mathbf{e}}'_k = \vec{\mathbf{n}}$ is fixed. An additional information is needed to fix the coordinate vectors $\vec{\mathbf{e}}'_i$ and $\vec{\mathbf{e}}'_j$ in the picture plane. An ,,upward" vector $\vec{\mathbf{u}}$ is chosen in the world coordinate system, and its projection into the picture plane defines the vertical direction i.e. the y'-direction of the picture system. The normed vector product of $\vec{\mathbf{u}}$ and $\vec{\mathbf{n}}$ defines the vector $\vec{\mathbf{e}}'_i$. Summary:

$$\vec{\mathbf{e}}'_k = \vec{\mathbf{n}}, \qquad \vec{\mathbf{e}}'_i = \frac{\vec{\mathbf{u}} \times \vec{\mathbf{n}}}{\|\vec{\mathbf{u}} \times \vec{\mathbf{n}}\|}, \qquad \vec{\mathbf{e}}'_j = \vec{\mathbf{e}}'_k \times \vec{\mathbf{e}}'_i\,. \tag{3.473}$$

The transformation matrices of the mapping from the world coordinate system into the picture system and their inverses are given as in (3.464) and (3.465) by substituting the corresponding coordinates of point $R(x_r, y_r, z_r)$ and vectors $\vec{\mathbf{e}}'_i$, $\vec{\mathbf{e}}'_j$ and $\vec{\mathbf{e}}'_k$.

3.5.5.3 Principal Projections

The projection is perpendicular to a plane which is perpendicular to one of the coordinate-axes. Depending on the direction of the projection and the viewing direction to the picture plane, the ground plan, the top view or one of the side views is formed on the projection plane.
The matrix form of the orthogonal projection of a point $P(x_\mathrm{P}, y_\mathrm{P}, z_\mathrm{P})$ to the projection plane being parallel to the x, y-plane with equation $z = z_0$ is

$$\begin{pmatrix} x'_\mathrm{P} \\ y'_\mathrm{P} \\ z'_\mathrm{P} \\ 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & z_0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_\mathrm{P} \\ y_\mathrm{P} \\ z_\mathrm{P} \\ 1 \end{pmatrix}. \tag{3.474}$$

In general, the projection plane is chosen $z_0 = 0$. The matrices of projections to the coordinate planes are

$$\mathbf{P}_x = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \qquad \mathbf{P}_y = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \qquad \mathbf{P}_z = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{3.475}$$

3.5.5.4 Axonometric Projection

In contrary to the orthographic projection, now the normal-vector $\vec{\mathbf{n}} = \{n_x, n_y, n_z\}$ of the projection plane and the direction of the projection are not parallel to any of the coordinate axes. There are three different cases to consider:

• **Isometric:** The angles of $\vec{\mathbf{n}}$ are the same with every coordinate axis. So, for the coordinates of $\vec{\mathbf{n}}$, $|n_x| = |n_y| = |n_z|$. The angle between the projected coordinate axes is 120°. Line segments being

parallel to the coordinate axes have the same distortion factor **(Abb.3.213a)**.

• **Dimetric:** $\vec{\mathbf{n}}$ has the same angle with two of the coordinate axes. In these directions equal distances remain equal. Two of the coordinates of $\vec{\mathbf{n}}$ have the same absolute value **(Fig.3.213b)**.

• **Trimetric:** $\vec{\mathbf{n}}$ has different angles with the coordinate axes. Consequently the coordinate axes have different distortion factors **(Fig.3.213c)**.

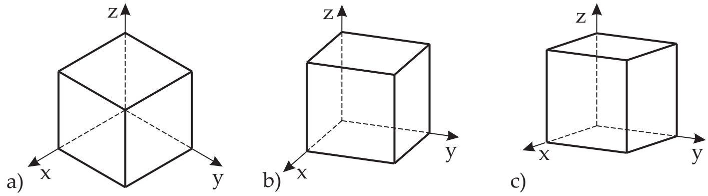


Figure 3.213

3.5.5.5 Isometric Projection

Consider the case when the projection plane contains the origin of the world coordinate system, and when its normal vector is $\vec{\mathbf{n}} = \left\{\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}\right\}$.

In order to determine the matrix of projection, a coordinate transformation into the projection coordinate system is combined with a subsequent projection along the z'-axis to the x', y'-plane. The definition of the picture coordinate system corresponds to (3.473). The upward vector is chosen as $\vec{\mathbf{u}} = \{0, 0, 1\}$. In this way the z-axis is mapped to the y'-axis. The unit basis-vectors of the picture coordinate system are:

$$\vec{\mathbf{e}}'_i = \frac{\vec{\mathbf{u}} \times \vec{\mathbf{n}}}{\|\vec{\mathbf{u}} \times \vec{\mathbf{n}}\|} = \left\{-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0\right\}, \qquad \vec{\mathbf{e}}'_j = \vec{\mathbf{e}}'_k \times \vec{\mathbf{e}}'_i = \left\{-\frac{1}{\sqrt{6}}, -\frac{1}{\sqrt{6}}, \frac{2}{\sqrt{6}}\right\},$$

$$\vec{\mathbf{e}}'_k = \vec{\mathbf{n}} = \left\{\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}\right\}.$$

The transformation matrices of mapping from the picture system into the world coordinate system and reverses with (3.464) and (3.465) are:

$$\overline{\mathbf{M}}_A = \begin{pmatrix} -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{3}} & 0 \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{3}} & 0 \\ 0 & \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{3}} & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \qquad \overline{\mathbf{M}}_A^{-1} = \overline{\mathbf{M}}_A^{\mathrm{T}} = \begin{pmatrix} -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 & 0 \\ -\frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} & 0 \\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{3.476}$$

Now, in the picture system the orthogonal projection along the z'-axis is:

$$\mathbf{P}_A = \mathbf{P}_z \overline{\mathbf{M}}_A^{-1} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 & 0 \\ -\frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} & 0 \\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 & 0 \\ -\frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{3.477}$$

The projection matrix $\mathbf{P}_A$ maps the points from the world coordinate system into the x', y'-plane of the picture system. By multiplication with matrix $\overline{M}_A$ one obtains the points in the world coordinate system. The projection-matrix of the complete projection is:

$$\mathbf{P} = \overline{\mathbf{M}}_A \mathbf{P}_z \overline{\mathbf{M}}_A^{-1} = \frac{1}{3}\begin{pmatrix} 2 & -1 & -1 & 0 \\ -1 & 2 & -1 & 0 \\ -1 & -1 & 2 & 0 \\ 0 & 0 & 0 & 3 \end{pmatrix}. \tag{3.478}$$

3.5.5.6 Oblique Parallel Projection

At an oblique projection the projecting rays intersect the projection plane in an angle β. In **Fig.3.214** the oblique projection of point $P(x_P, y_P, z_P)$ is $P'(x'_P, y'_P, z'_P)$ and its orthogonal projection is P'_0. L is the length of the projected line segment $\overline{P'_0 P'}$. The projection is characterized by two quantities. $d = \dfrac{L}{z} = \dfrac{1}{\tan\beta}$ gives the factor by which the line segment perpendicular to the projection plane is scaled. α is the angle between the x-axis and the projected image of the perpendicular line segment. Then the rule for the coordinates of the whole projection is:

$$x'_P = x_P - z_P\, d\cos\alpha\,, \quad y'_P = y_P - z_P\, d\sin\alpha\,, \quad z'_P = 0 \qquad \text{or} \tag{3.479}$$

$$\begin{pmatrix} x'_P \\ y'_P \\ z'_P \\ 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & -d\cos\alpha & 0 \\ 0 & 1 & -d\sin\alpha & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_P \\ y_P \\ z_P \\ 1 \end{pmatrix}. \tag{3.480}$$

If the projection plane is different from the x, y-plane, then a coordinate transformation into the picture system must be done before applying the projection-matrix (see 3.5.5.2, p. 238).

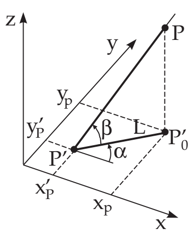


Figure 3.214

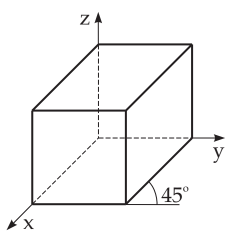


Figure 3.215

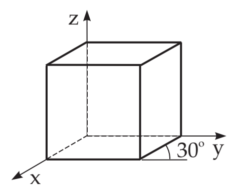


Figure 3.216

■ Cavalier projection of the unit cube with $\alpha = 45°$ and $d = 1$, i.e. $\beta = 45°$. Segments being perpendicular to the projection plane do not become shorter in this case **(Fig.3.215)**.
The four vertices not lying in the x, y-plane are put as columns into a matrix. This matrix multiplied by the matrix of the projection will give the coordinates of the projected points:

$$\begin{pmatrix} -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 1-\frac{1}{\sqrt{2}} & 1-\frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & 1-\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 1-\frac{1}{\sqrt{2}} \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & -\frac{1}{\sqrt{2}} & 0 \\ 0 & 1 & -\frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix}.$$

■ Cabinet projection of the unit cube with $\alpha = 30°$, $d = 1/2$, i.e. $\beta = 63.4°$. The lengths of segments perpendicular to the projection plane are halved **(Fig.3.216)**.

The coordinates of the projections of the four vertices not lying in the x, y-plane are calculated by

$$\begin{pmatrix} -\frac{\sqrt{3}}{4} & 1-\frac{\sqrt{3}}{4} & -\frac{\sqrt{3}}{4} & 1-\frac{\sqrt{3}}{4} \\ -\frac{1}{4} & -\frac{1}{4} & \frac{3}{4} & \frac{3}{4} \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & -\frac{\sqrt{3}}{4} & 0 \\ 0 & 1 & -\frac{1}{4} & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix}.$$

3.5.5.7 Perspective Projection

1. Mapping Formulation

The formulation of the mapping of a perspective projection can be given reasonably in the picture coordinate system. The origin is chosen so that the center of projection lies on the z-axis.

As it can be seen in **Fig.3.217**, the relations between the coordinates of a point $P(x_P, y_P, z_P)$ and its projected image $P'(x'_P, y'_P, z'_P)$ can be calculated using equations of the intercept theorems (see 3.1.3.2,**5.**, p. 135):

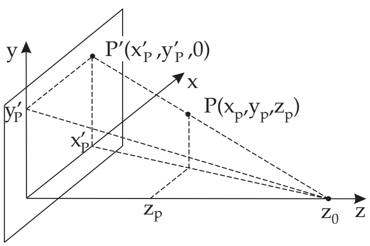


Figure 3.217

$$\frac{x'_P}{x_P} = \frac{z_0}{z_0 - z_P} = \frac{1}{1 - \frac{z_P}{z_0}}, \quad \frac{y'_P}{y_P} = \frac{1}{1 - \frac{z_P}{z_0}},$$

$$z'_P = 0.$$

The relations between the original and the image coordinates are not linear. However using the properties of homogeneous coordinates (see 3.5.4.2, p. 231) the projection rule can be given in the following matrix form:

$$\begin{pmatrix} x'_P \\ y'_P \\ z'_P \\ 1 \end{pmatrix} = \begin{pmatrix} x_P \\ y_P \\ 0 \\ 1 - \frac{z_P}{z_0} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -\frac{1}{z_0} & 1 \end{pmatrix} \begin{pmatrix} x_P \\ y_P \\ z_P \\ 1 \end{pmatrix}. \tag{3.481}$$

2. Vanishing points

Perspective projections have the property, that parallel lines, which are not parallel to the picture plane, seem to intersect each other in a point. This point is called *vanishing point*. The vanishing points of lines being parallel to the coordinate axes are called *principal points* or *principal vanishing points*. The number of principal points is equal to the number of coordinate axes intersecting the picture plane. In **Fig.3.218** there are perspective representations with one, two and three principal points. The

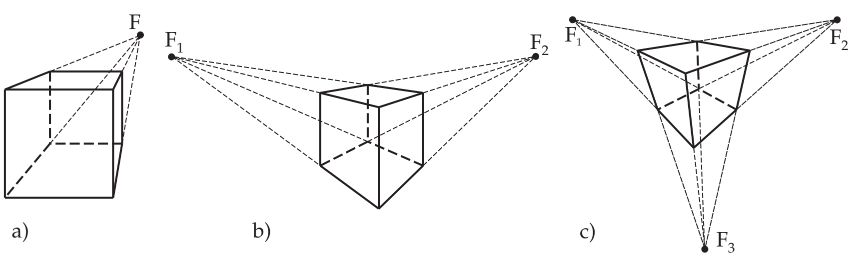


Figure 3.218

principal points coincide with the intersection points of the picture plane and the rays being parallel to the coordinate axes. If the center of projection is $C(x_c, y_c, z_c)$, a point of the picture plane is $R(x_r, y_r, z_r)$ and the normal vector is $\vec{\mathbf{n}} = \{n_x, n_y, n_z\}$, then the coordinates of the principal points are:

$$F_x(x_c + \frac{d}{n_x}, y_c, z_c), \quad F_y(x_c, y_c + \frac{d}{n_y}, z_c), \quad F_z(x_c, y_c, z_c + \frac{d}{n_z}) \quad \text{where} \tag{3.482a}$$

$$d = (\vec{\mathbf{c}} - \vec{\mathbf{r}}) \cdot \vec{\mathbf{n}} = (x_c - x_r)n_x + (y_c - y_r)n_y + (z_c - z_r)n_z \tag{3.482b}$$

in the case they exist. If a coordinate of the normal vector $\vec{\mathbf{n}}$ is zero, then there is no principal point in that direction.

3.6 Differential Geometry

In differential geometry planar curves and curves and surfaces in space are discussed by the methods of differential calculus. Therefore it is to be supposed that the functions describing the curves and surfaces are continuous and continuously differentiable as many times as necessary for discussion of the corresponding properties. The absence of these assumptions is allowed only at a few points of the curves and surfaces. These points are called *singular points*.

During the discussion of geometric configurations with their equations there are to be distinguished properties depending on the choice of the coordinate system, such as intersection points with the coordinate axes, the slope or direction of tangent lines, maxima, minima; and invariant properties independent of coordinate transformations, such as inflection points, curvature, and cyclic points. There are also local properties, which are valid only for a small part of the curves and surfaces as the curvature and differential of arc or area of surfaces, and there are properties belonging to the whole curve or surface, such as number of vertices, and arc length of a closed curve.

3.6.1 Plane Curves

3.6.1.1 Ways to Define a Plane Curve

1. Coordinate Equations

A plane curve can be analytically defined in the following ways.

1. In Cartesian coordinates:

a) Implicit: $F(x, y) = 0,$ (3.483)

b) Explicit: $y = f(x),$ (3.484)

2. In Parametric Form: $x = x(t), \quad y = y(t).$ (3.485)

3. In Polar Coordinates: $\rho = f(\varphi).$ (3.486)

2. Positive Direction on a Curve

If a curve is given in the form (3.485), the positive direction is defined on it in which a point $P\,(x(t), y(t))$ of the curve moves for increasing values of the parameter t. If the curve is given in the form (3.484), then the abscissa can be considered as a parameter $(x = x, y = f(x))$, so the positive direction holds for increasing abscissa. For the form (3.486) the angle can be considered as a parameter φ $(x = f(\varphi)\cos\varphi$, $y = f(\varphi)\sin\varphi)$, so the positive direction holds for increasing φ, i.e., counterclockwise.

■ **Fig.3.219a, b, c**: **A:** $x = t^2$, $y = t^3$, **B:** $y = \sin x$, **C:** $\rho = a\varphi$.

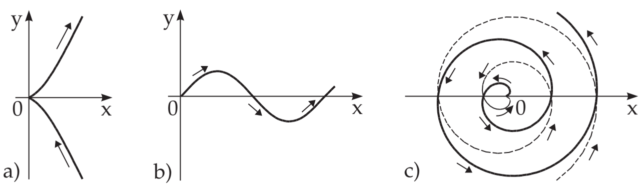


Figure 3.219

3.6.1.2 Local Elements of a Curve

Depending on whether a changing point P on the curve is given in the form (3.484), (3.485) or (3.486), its position is defined by x, t or φ. A point arbitrarily close to P with parameter values $x + dx$, $t + dt$ or $\varphi + d\varphi$ is denoted here by N.

1. Differential of Arc

If s denotes the length of the curve from a fixed point A to the point P, the infinitesimal increment $\Delta s = \overset{\frown}{PN}$ can be expressed approximately by the differential ds of the arc length, the *differential of*

arc:

$$\Delta s \approx ds = \begin{cases} \sqrt{1+\left(\dfrac{dy}{dx}\right)^2}\,dx & \text{for the form (3.484),} \quad (3.487)\\ \sqrt{x'^2+y'^2}\,dt & \text{for the form (3.485),} \quad (3.488)\\ \sqrt{\rho^2+\rho'^2}\,d\varphi & \text{for the form (3.486).} \quad (3.489)\end{cases}$$

■ **A:** $y = \sin x,\ ds = \sqrt{1+\cos^2 x}\,dx$. ■ **B:** $x = t^2,\ y = t^3,\ ds = t\sqrt{4+9t^2}\,dt$.
■ **C:** $\rho = a\varphi\ \ (a > 0), \quad ds = a\sqrt{1+\varphi^2}\,d\varphi$.

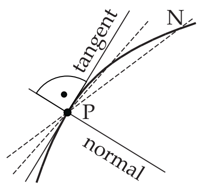


Figure 3.220

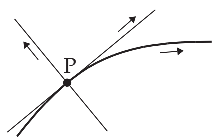


Figure 3.221

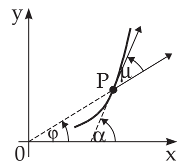


Figure 3.222

2. Tangent and Normal

1. Tangent at a Point P to a Curve is a line in the limiting position of the secants PN for $N \to P$; the *normal* is a line through P which is perpendicular to the tangent here **(Fig. 3.220)**.

2. The Equations of the Tangent and the Normal are given in **Table 3.26** for the three cases (3.483), (3.484), and (3.485). Here x, y are the coordinates of P, and X, Y are the coordinates of the points of the tangent and normal. The values of the derivatives should be calculated at the point P.

Table 3.26 Tangent and normal equations

Type of equation	Equation of the tangent	Equation of the normal
(3.483)	$\dfrac{\partial F}{\partial x}(X-x)+\dfrac{\partial F}{\partial y}(Y-y)=0$	$\dfrac{X-x}{\dfrac{\partial F}{\partial x}}=\dfrac{Y-y}{\dfrac{\partial F}{\partial y}}$
(3.484)	$Y-y=\dfrac{dy}{dx}(X-x)$	$Y-y=-\dfrac{1}{\dfrac{dy}{dx}}(X-x)$
(3.485)	$\dfrac{Y-y}{y'}=\dfrac{X-x}{x'}$	$x'(X-x)+y'(Y-y)=0$

Examples for equations of the tangent and normal for the following curves:

■ **A:** Circle $x^2+y^2=25$ at the point $P(3,4)$:
a) Equation of the tangent: $2x(X-x)+2y(Y-y)=0$ or $Xx+Yy=25$ considering that the point P lies on the circle: $3X+4Y=25$.
b) Equation of the normal: $\dfrac{X-x}{2x}=\dfrac{Y-y}{2y}$ or $Y=\dfrac{y}{x}X$; at the point P: $Y=\dfrac{4}{3}X$.

■ **B:** Sine curve $y=\sin x$ at the point $0(0,0)$:
a) Equation of the tangent: $Y-\sin x=\cos x(X-x)$ or $Y=X\cos x+\sin x-x\cos x$; at the point $(0,0)$: $Y=X$.

b) Equation of the normal: $Y - \sin x = -\frac{1}{\cos x}(X - x)$ or $Y = -X \sec x + \sin x + x \sec x$; at the point $(0,0)$: $Y = -X$.

■ **C:** Curve with $x = t^2, y = t^3$ at the point $P(4,-8), t = -2$:

a) Equation of the tangent: $\frac{Y - t^3}{3t^2} = \frac{X - t^2}{2t}$ or $Y = \frac{3}{2}tX - \frac{1}{2}t^3$; at the point P: $Y = -3X + 4$.

b) Equation of the normal: $2t\,(X - t^2) + 3t^2\,(Y - t^3) = 0$ or $2X + 3tY = t^2\,(2 + 3t^2)$; at the point $P(4,-8)$: $X - 3Y = 28$.

3. Positive Direction of the Tangent and Normal of the Curve If the curve is given in one of the forms (3.484), (3.485), (3.486), the positive directions on the tangent and normal are defined in the following way: The positive direction of the tangent is the same as on the curve at the point of contact, and one gets the positive direction on the normal from the positive direction of the tangent by rotating it counterclockwise around P by an angle of 90° **(Fig. 3.221)**. The tangent and the normal are divided into a positive and a negative half-line by the point P.

4. The Slope of the Tangent can be determined

a) by the *angle of slope of the tangent* α, between the positive directions of the axis of abscissae and the tangent, or

b) if the curve is given in polar coordinates, by the angle μ, between the radius vector OP ($\overline{OP} = \rho$) and the positive direction of the tangent **(Fig. 3.222)**. For the angles α and μ the following formulas are valid, where ds is calculated according to (3.487)–(3.489):

$$\tan\alpha = \frac{dy}{dx}, \quad \cos\alpha = \frac{dx}{ds}, \quad \sin\alpha = \frac{dy}{ds}; \tag{3.490a}$$

$$\tan\mu = \frac{\rho}{\frac{d\rho}{d\varphi}}, \quad \cos\mu = \frac{d\rho}{ds}, \quad \sin\mu = \rho\frac{d\varphi}{ds}. \tag{3.490b}$$

■ **A:** $y = \sin x$, $\quad \tan\alpha = \cos x$, $\quad \cos\alpha = \frac{1}{\sqrt{1+\cos^2 x}}$, $\quad \sin\alpha = \frac{\cos x}{\sqrt{1+\cos^2 x}}$;

■ **B:** $x = t^2,\ y = t^3$, $\quad \tan\alpha = \frac{3t}{2}$, $\quad \cos\alpha = \frac{2}{\sqrt{4+9t^2}}$, $\quad \sin\alpha = \frac{3t}{\sqrt{4+9t^2}}$;

■ **C:** $\rho = a\varphi$, $\quad \tan\mu = \varphi$, $\quad \cos\mu = \frac{1}{\sqrt{1+\varphi^2}}$, $\quad \sin\mu = \frac{\varphi}{\sqrt{1+\varphi^2}}$.

5. Segments of the Tangent and Normal, Subtangent and Subnormal (Fig. 3.223)

a) In Cartesian Coordinates for the definitions in form (3.484), (3.485):

$$\overline{PT} = \left|\frac{y}{y'}\sqrt{1+y'^2}\right| \quad \textit{(segment of the tangent)}, \tag{3.491a}$$

$$\overline{PN} = \left|y\sqrt{1+y'^2}\right| \quad \textit{(segment of the normal)}, \tag{3.491b}$$

$$\overline{P'T} = \left|\frac{y}{y'}\right| \quad \textit{(subtangent)}, \tag{3.491c}$$

$$\overline{P'N} = |yy'| \quad \textit{(subnormal)}. \tag{3.491d}$$

b) In Polar Coordinates for the definitions in form (3.486):

$$\overline{PT'} = \left|\frac{\rho}{\rho'}\sqrt{\rho^2+\rho'^2}\right| \quad \textit{(segment of the polar tangent)}, \tag{3.492a}$$

$$\overline{PN'} = \left|\sqrt{\rho^2+\rho'^2}\right| \quad \textit{(segment of the polar normal)}, \tag{3.492b}$$

$$\overline{OT'} = \left|\frac{\rho^2}{\rho'}\right| \quad (\textit{polar subtangent}), \text{(3.492c)} \qquad \overline{ON'} = |\rho'| \quad (\textit{polar subnormal}). \qquad \text{(3.492d)}$$

■ **A:** $y = \cosh x,\ y' = \sinh x,\ \sqrt{1+y'^2} = \cosh x;\ \overline{PT} = |\cosh x \coth x|,\ \overline{PN} = |\cosh^2 x|,\ \overline{P'T} = |\coth x|,\ \overline{P'N} = |\sinh x \cosh x|.$

■ **B:** $\rho = a\varphi\ (a > 0),\ \rho' = a,\ \sqrt{\rho^2+\rho'^2} = a\sqrt{1+\varphi^2};\ \overline{PT'} = \left|a\varphi\sqrt{1+\varphi^2}\right|,\ \overline{PN'} = \left|a\sqrt{1+\varphi^2}\right|,\ \overline{OT'} = \left|a\varphi^2\right|,\ \overline{ON'} = a.$

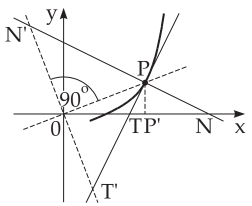


Figure 3.223

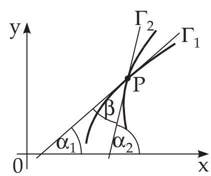


Figure 3.224

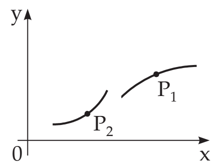


Figure 3.225

6. Angle Between Two Curves The *angle β between two curves Γ_1* and Γ_2 at their intersection point P is defined as the angle between their tangents at the point P **(Fig. 3.224)**. By this definition the calculation of the angle β has been reduced to the calculation of the angle between two lines with slopes

$$k_1 = \tan\alpha_1 = \left(\frac{df_1}{dx}\right)_P, \qquad \text{(3.493a)} \qquad\qquad k_2 = \tan\alpha_2 = \left(\frac{df_2}{dx}\right)_P. \qquad \text{(3.493b)}$$

Here $y = f_1(x)$ is the equation of Γ_1 and $y = f_2(x)$ is the equation of Γ_2, and the derivatives are calculated at the point P. One gets β with the help of the formula

$$\tan\beta = \tan(\alpha_1 - \alpha_2) = \frac{\tan\alpha_2 - \tan\alpha_1}{1 + \tan\alpha_1 \tan\alpha_2}. \qquad \text{(3.494)}$$

■ Determine the angle between the parabolas $y = \sqrt{x}$ and $y = x^2$ at the point $P(1,1)$:

$$\tan\alpha_1 = \left(\frac{d\sqrt{x}}{dx}\right)_{x=1} = \frac{1}{2}, \quad \tan\alpha_2 = \left(\frac{d\,(x^2)}{dx}\right)_{x=1} = 2, \quad \tan\beta = \frac{\tan\alpha_2 - \tan\alpha_1}{1 + \tan\alpha_1\tan\alpha_2} = \frac{3}{4}.$$

3. Convex and Concave Part of a Curve

If a curve is given in the explicit form $y = f(x)$, then a small part containing the point P can be examined if the curve is concave up or down here, except of course if P is an inflection point or a singular point (see 3.6.1.3, p. 249). If the second derivative $f''(x) > 0$ (if it exists), then the curve is concave up, i.e., in the direction of positive y (point P_2 in **Fig. 3.225)**. If $f''(x) < 0$ holds (point P_1), then the curve is concave down. In the case if $f''(x) = 0$ holds, it should be checked if it is an inflection point.

■ $y = x^3$ **(Fig. 2.15b)**; $y'' = 6x$, for $x > 0$ the curve is concave up, for $x < 0$ concave down.

4. Curvature and Radius of Curvature

1. Curvature of a Curve The curvature K of a curve at the point P is the limit of the ratio of the angle δ between the positive tangent directions at the points P and N **(Fig. 3.226)** and the arc length

$\overset{\frown}{PN}$ for $\overset{\frown}{PN} \to 0$:

$$K = \lim_{\overset{\frown}{PN} \to 0} \frac{\delta}{\overset{\frown}{PN}}. \tag{3.495}$$

The sign of the curvature K depends on whether the curve bends toward the positive half of the normal ($K > 0$) or toward the negative half of it ($K < 0$) (see 3.6.1.1, **2.**, p. 245). In other words the center of curvature for $K > 0$ is on the positive side of the normal, for $K < 0$ it is on the negative side. Sometimes the curvature K is considered only as a positive quantity. Then it is to take the absolute value of the limit above.

2. Radius of Curvature of a Curve The radius of curvature R of a curve at the point P is the reciprocal value of the absolute value of the curvature:

$$R = \frac{1}{|K|}. \tag{3.496}$$

The larger the curvature K is at a point P the smaller the radius of curvature R is.

■ **A:** For a circle with radius a the curvature $K = 1/a$ and the radius of curvature $R = a$ are constant for every point.

■ **B:** For a line with $K = 0$ holds $R = \infty$.

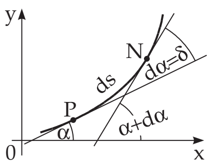


Figure 3.226

3. Formulas for Curvature and Radius of Curvature With the notation $\delta = d\alpha$ and $\overset{\frown}{PN} = ds$ **(Fig. 3.226)** in general

$$K = \frac{d\alpha}{ds}, \quad R = \left|\frac{ds}{d\alpha}\right|. \tag{3.497}$$

For the different defining formulas of curves in 3.6.1.1, p. 243 the different expressions for K and R are:

Definition as in (3.484): $$K = \frac{\dfrac{d^2y}{dx^2}}{\left[1 + \left(\dfrac{dy}{dx}\right)^2\right]^{3/2}}, \qquad R = \left|\frac{\left[1 + \left(\dfrac{dy}{dx}\right)^2\right]^{3/2}}{\dfrac{d^2y}{dx^2}}\right|, \tag{3.498}$$

Definition as in (3.485): $$K = \frac{\begin{vmatrix} x' & y' \\ x'' & y'' \end{vmatrix}}{\left(x'^2 + y'^2\right)^{3/2}}, \qquad R = \left|\frac{\left(x'^2 + y'^2\right)^{3/2}}{\begin{vmatrix} x' & y' \\ x'' & y'' \end{vmatrix}}\right|, \tag{3.499}$$

Definition as in (3.483): $$K = \frac{\begin{vmatrix} F_{xx} & F_{xy} & F_x \\ F_{yx} & F_{yy} & F_y \\ F_x & F_y & 0 \end{vmatrix}}{\left({F_x}^2 + {F_y}^2\right)^{3/2}}, \qquad R = \left|\frac{\left({F_x}^2 + {F_y}^2\right)^{3/2}}{\begin{vmatrix} F_{xx} & F_{xy} & F_x \\ F_{yx} & F_{yy} & F_y \\ F_x & F_y & 0 \end{vmatrix}}\right|, \tag{3.500}$$

Definition as in (3.486): $$K = \frac{\rho^2 + 2\rho'^2 - \rho\rho''}{(\rho^2 + \rho'^2)^{3/2}}, \qquad R = \left|\frac{(\rho^2 + \rho'^2)^{3/2}}{\rho^2 + 2\rho'^2 - \rho\rho''}\right|. \tag{3.501}$$

■ **A:** $y = \cosh x, \quad K = \dfrac{1}{\cosh^2 x}$; ■ **B:** $x = t^2, \quad y = t^3, \quad K = \dfrac{6}{t(4+9t^2)^{3/2}}$;

■ **C:** $y^2 - x^2 = a^2, \quad K = \dfrac{a^2}{(x^2+y^2)^{3/2}}$; ■ **D:** $\rho = a\varphi, \quad K = \dfrac{1}{a} \cdot \dfrac{\varphi^2+2}{(\varphi^2+1)^{3/2}}$.

5. Circle of Curvature and Center of Curvature

1. Circle of Curvature at the point P is the limiting position of the circles passing through P and two points N and M of the curve from its neighborhood, for $N \to P$ and $M \to P$ **(Fig. 3.227)**. It goes through the point of the curve and here it has the same first and the same second derivative as the curve. Therefore it fits the curve at the point of contact especially well. It is also called the *osculating circle*. Its radius is the *radius of curvature*. It is obvious that it is the reciprocal value of the absolute value of the curvature.

2. Center of Curvature The center C of the circle of curvature is the center of curvature of the point P. It is on the concave side of the curve, and on the normal of the curve.

3. Coordinates of the Center of Curvature The coordinates (x_C, y_C) of the center of curvature for curves with defining equations in 3.6.1.1, p. 243 can be determined from the following formulas.

Definition as in (3.484): $$x_C = x - \frac{\dfrac{dy}{dx}\left[1+\left(\dfrac{dy}{dx}\right)^2\right]}{\dfrac{d^2y}{dx^2}}, \qquad y_C = y + \frac{1+\left(\dfrac{dy}{dx}\right)^2}{\dfrac{d^2y}{dx^2}}. \tag{3.502}$$

Definition as in (3.485): $$x_C = x - \frac{y'(x'^2+y'^2)}{\begin{vmatrix} x' & y' \\ x'' & y'' \end{vmatrix}}, \qquad y_C = y + \frac{x'(x'^2+y'^2)}{\begin{vmatrix} x' & y' \\ x'' & y'' \end{vmatrix}}. \tag{3.503}$$

Definition as in (3.486): $$x_C = \rho\cos\varphi - \frac{(\rho^2+\rho'^2)(\rho\cos\varphi + \rho'\sin\varphi)}{\rho^2+2\rho'^2-\rho\rho''},$$
$$y_C = \rho\sin\varphi - \frac{(\rho^2+\rho'^2)(\rho\sin\varphi - \rho'\cos\varphi)}{\rho^2+2\rho'^2-\rho\rho''}. \tag{3.504}$$

Definition as in (3.483): $$x_C = x + \frac{F_x\left(F_x^2+F_y^2\right)}{\begin{vmatrix} F_{xx} & F_{xy} & F_x \\ F_{yx} & F_{yy} & F_y \\ F_x & F_y & 0 \end{vmatrix}}, \qquad y_C = y + \frac{F_y\left(F_x^2+F_y^2\right)}{\begin{vmatrix} F_{xx} & F_{xy} & F_x \\ F_{yx} & F_{yy} & F_y \\ F_x & F_y & 0 \end{vmatrix}}. \tag{3.505}$$

These formulas can be transformed into the form
$$x_C = x - R\sin\alpha, \quad y_C = y + R\cos\alpha \quad \text{or} \tag{3.506}$$
$$x_C = x - R\frac{dy}{ds}, \quad y_C = y + R\frac{dx}{ds} \tag{3.507}$$
(Fig. 3.228), where R should be calculated as in (3.498)–(3.501).

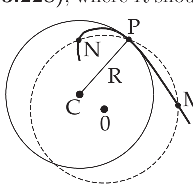


Figure 3.227

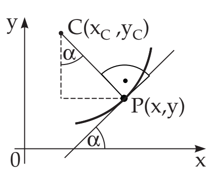


Figure 3.228


Figure 3.229

3.6.1.3 Special Points of a Curve

In the following are to be discussed only the points remaining invariant during coordinate transformations. In order to determine maxima and minima see 6.1.5.3, p. 444.

1. Inflection Points and the Rules to Determine Them

Inflection points are the points of the curve where the curvature changes its sign **(Fig. 3.229)** while a tangent exists. The tangent line at the inflection point intersects the curve, so the curve is on both sides of the line in this neighborhood. At the inflection point $K = 0$ and $R = \infty$ hold.

1. Explicit Form (3.484) of the Curve $y = f(x)$

a) A Necessary Condition for the existence of an inflection point is the zero value of the second derivative

$$f''(x) = 0 \tag{3.508}$$

if it exists at the inflection point (for the case of non-existent second derivative see **b)**). In order to determine the inflection points for existing second derivative all the roots of the equation $f''(x) = 0$ with values $x_1, x_2, \ldots, x_i, \ldots, x_n$ are to be considered, and to be substituted into the further derivatives. If for a value x_i the first non-zero derivative has odd order, then there is an inflection point here. If the considered point is not an inflection point, because for the first non-disappearing derivative of kth order, k is an even number, then for $f^{(k)}(x) < 0$ the curve is concave down; for $f^{(k)}(x) > 0$ it is concave up. If the higher-order derivatives are not be checked, for instance in the case they do not exist, see point **b)**.

b) A Sufficient Condition for the existence of an inflection point is the change of the sign of the second derivative $f''(x)$ while traversing from the left neighborhood of this point to the right, if also a tangent exists here, of course. So the question, of whether the curve has an inflection point at the point with abscissa x_i, can be answered by checking the sign of the second derivative traversing the considered point: If the sign changes during the traverse, there is an inflection point. (Since x_i is a root of the second derivative, the function has a first derivative, and consequently the curve has a tangent.) This method can also be used in the case if $y'' = \infty$, together with the checking of the existence of a tangent line, e.g. in the case of a vertical tangent line.

■ **A:** $y = \dfrac{1}{1+x^2}$, $f''(x) = -2\dfrac{1-3x^2}{(1+x^2)^3}$, $x_{1,2} = \pm\dfrac{1}{\sqrt{3}}$, $f'''(x) = 24x\dfrac{1-x^2}{(1+x^2)^4}$, $f'''(x_{1,2}) \neq 0$.

Inflection points: $A\left(\dfrac{1}{\sqrt{3}}, \dfrac{3}{4}\right)$, $B\left(-\dfrac{1}{\sqrt{3}}, \dfrac{3}{4}\right)$.

■ **B:** $y = x^4$, $f''(x) = 12x^2$, $x_1 = 0$, $f'''(x) = 24x$, $f'''(x_1) = 0$, $f^{IV}(x) = 24$; there is no inflection point.

■ **C:** $y = x^{\frac{5}{3}}$, $y' = \dfrac{5}{3}x^{\frac{2}{3}}$, $y'' = \dfrac{10}{9}x^{-\frac{1}{3}}$; for $x = 0$ we have $y'' = \infty$.

As the value of x changes from negative to positive, the second derivative changes its sign from "−" to "+", so the curve has an inflection point at $x = 0$.

Remark: In practice, if from the shape of the curve the existence of inflection points follows, for instance between a minimum and a maximum with continuous derivatives, then there are to be determined only the points x_i without checking the further derivatives.

2. Other Defining Forms of the Curve The necessary condition (3.508) for the existence of an inflection point in the case of the defining form of the curve (3.484) will have the analytic form for the other defining formulas as follows:

a) Definition in parametric form as in (3.485):

$$\begin{vmatrix} x' & y' \\ x'' & y'' \end{vmatrix} = 0. \tag{3.509}$$

b) Definition in polar coordinates as in (3.486):

$$\rho^2 + 2\rho'^2 - \rho\rho'' = 0. \tag{3.510}$$

c) Definition in implicit form as in (3.483): $F(x,y)=0$ and $\begin{vmatrix} F_{xx} & F_{xy} & F_x \\ F_{yx} & F_{yy} & F_y \\ F_x & F_y & 0 \end{vmatrix} = 0.$ (3.511)

In these cases the system of solutions gives the possible coordinates of inflection points.

■ **A:** $x = a\left(t - \frac{1}{2}\sin t\right), \quad y = a\left(1 - \frac{1}{2}\cos t\right)$ (curtated cycloid **(Fig. 2.68b)**, p. 102);

$\begin{vmatrix} x' & y' \\ x'' & y'' \end{vmatrix} = \frac{a^2}{4}\begin{vmatrix} 2-\cos t & \sin t \\ \sin t & \cos t \end{vmatrix} = \frac{a^2}{4}(2\cos t - 1); \quad \cos t_k = \frac{1}{2}; \quad t_k = \pm\frac{\pi}{3} + 2k\pi \quad (k = 0, \pm 1, \pm 2, \ldots).$
The curve has an infinite number of inflection points for the parameter values t_k.

■ **B:** $\rho = \frac{1}{\sqrt{\varphi}}$; $\rho^2 + 2\rho'^2 - \rho\rho'' = \frac{1}{\varphi} + \frac{1}{2\varphi^3} - \frac{3}{4\varphi^3} = \frac{1}{4\varphi^3}(4\varphi^2 - 1)$. The inflection point is at the angle $\varphi = 1/2$.

■ **C:** $x^2 - y^2 = a^2$ (hyperbola). $\begin{vmatrix} F_{xx} & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \end{vmatrix} = \begin{vmatrix} 2 & 0 & 2x \\ 0 & -2 & -2y \\ 2x & -2y & 0 \end{vmatrix} = 8x^2 - 8y^2$. The equations $x^2 - y^2 = a^2$ and $8(x^2 - y^2) = 0$ contradict each other, so the hyperbola has no inflection point.

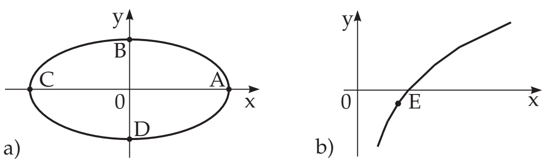


Figure 3.230

2. Vertices

Vertices are the points of the curve where the curvature has a maximum or a minimum. The ellipse has for instance four vertices A, B, C, D, the curve of the logarithm function has one vertex at $E\,(1/\sqrt{2}\,, -\ln 2/2)$ **(Fig. 3.230)**. The determination of vertices is reduced to the determination of the extreme values of K or, if it is simpler, the extreme values of R. The formulas (3.498)–(3.501) can be used for the calculations.

3. Singular Points

Singular point is a general notion for different special points of a curve.

1. Types of Singular Points The points a), b), etc. to j) correspond to the representation in **Fig. 3.231**.

a) Double Point: At a *double point* the curve intersects itself **(Fig. 3.231a)**.

b) Isolated Point: An *isolated point* satisfies the equation of the curve; but it is separated from the curve **(Fig. 3.231b)**.

c), d) Cuspidal point: At a *cuspidal point* or briefly a *cusp* the orientation of the curve changes; according to the position of the tangent a cusp of the first kind and a cusp of the second kind **(Fig.3.231c,d)** can be distinguished.

e) Tacnode or point of osculation: At the *tacnode* the curve contacts itself **(Fig. 3.231e)**.

f) Corner point: At a *corner point* the curve suddenly changes its direction but in contrast to a cusp there are two different tangents for the two different branches of the curve here **(Fig.3.231f)**.

g) Terminal point: At a *terminal point* the curve terminates **(Fig.3.231g)**.

h) Asymptotic point: In the neighborhood of an *asymptotic point* the curve usually winds in and out

or around infinitely many times, while it approaches itself and the point arbitrarily close **(Fig.3.231h)**.

i), j) More Singularities: It is possible that the curve has two or more such singularities at the same point **(Fig. 3.231i,j)**.

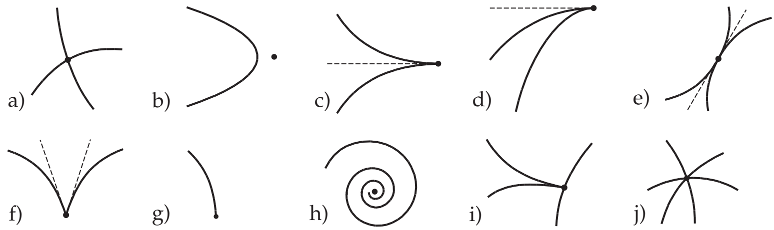


Figure 3.231

2. Determination of the Tacnode, Corner, Terminal, and Asymptotic Points Singularities of these types occur only on the curves of transcendental functions (see 3.5.2.5, p. 195).
The corner point corresponds to a finite jump of the derivative dy/dx.
Points where the function terminates correspond to the points of discontinuity of the function $y = f(x)$ with a finite jump or to a direct termination.
Asymptotic points can be determined in the easiest way in the case of curves given in polar coordinates as $\rho = f(\varphi)$. If for $\varphi \to \infty$ or $\varphi \to -\infty$ the limit is equal to zero ($\lim \rho = 0$), i.e., the pole is an asymptotic point.

■ **A:** The origin is a corner point for the curve $y = \dfrac{x}{1 + e^{\frac{1}{x}}}$ **(Fig. 6.2c)** (see 6.1.1, p. 432).

■ **B:** The points $(1, 0)$ and $(1, 1)$ are points of discontinuity of the function $y = \dfrac{1}{1 + e^{\frac{1}{x-1}}}$ **(Fig. 2.8)** (see 2.1.4.5, p. 106).

■ **C:** The logarithmic spiral $\rho = ae^{k\varphi}$ **(Fig. 2.75)** (see 2.14.3, p. 106) has an asymptotic point at the origin.

3. Determination of Multiple Points (Cases from a) to e), and i), and j)) Double points, triple points, etc. are denoted by the general term *multiple points*. To determine them, one starts with the equation of the curve of the form $F(x, y) = 0$. A point A with coordinates (x_1, y_1) satisfying the three equations $F = 0$, $F_x = 0$, and $F_y = 0$ is a double point if at least one of the three derivatives of second order F_{xx}, F_{xy}, and F_{yy} does not vanish. Otherwise A is a triple point or a point with higher multiplicity.
The properties of a double point depend on the sign of the Jacobian determinant

$$\Delta = \begin{vmatrix} F_{xx} & F_{xy} \\ F_{yx} & F_{yy} \end{vmatrix}_{\begin{pmatrix} x=x_1 \\ y=y_1 \end{pmatrix}} . \tag{3.512}$$

Case $\Delta < 0$: For $\Delta < 0$ the curve intersects itself at the point A; the slopes of the tangents at A are the roots of the equation

$$F_{yy}k^2 + 2F_{xy}k + F_{xx} = 0. \tag{3.513}$$

Case $\Delta > 0$: For $\Delta > 0$ A is an isolated point.
Case $\Delta = 0$: For $\Delta = 0$ A is either a cusp or a tacnode; the slope of the tangent is

$$\tan\alpha = -\frac{F_{xy}}{F_{yy}} . \tag{3.514}$$

For more precise investigation about multiple points the origin can be translated to the point A, and rotate so that the x-axis becomes a tangent at A. Then from the form of the equation one can tell if it is a cusp of first or second kind, or if it is a tacnode.

■ **A:** $F(x,y) \equiv (x^2+y^2)^2 - 2a^2(x^2-y^2) = 0$ (Lemniscate, **Fig. 2.66**, 2.12.6, p. 101); $F_x = 4x(x^2+y^2-a^2)$, $F_y = 4y(x^2+y^2+a^2)$; the equation system $F_x = 0$, $F_y = 0$ results in the three solutions $(0,0)$, $(\pm a, 0)$, from which only the first one satisfies the condition $F = 0$. Substituting $(0,0)$ into the second derivatives yields $F_{xx} = -4a^2$, $F_{xy} = 0$, $F_{yy} = +4a^2$; $\Delta = -16a^4 < 0$, i.e., at the origin the curve intersects itself; the slopes of the tangents are $\tan\alpha = \pm 1$, their equations are $y = \pm x$.

■ **B:** $F(x,y) \equiv x^3 + y^3 - x^2 - y^2 = 0$; $F_x = x(3x-2)$, $F_y = y(3y-2)$; among the points $(0,0)$, $(0,2/3)$, $(2/3,0)$, and $(2/3,2/3)$ only the first one belongs to the curve; there is $F_{xx} = -2$, $F_{xy} = 0$, $F_{yy} = -2$, $\Delta = 4 > 0$, i.e., the origin is an isolated point.

■ **C:** $F(x,y) \equiv (y-x^2)^2 - x^5 = 0$. The equations $F_x = 0$, $F_y = 0$ result in only one solution $(0,0)$, it also satisfies the equation $F = 0$. Furthermore there is $\Delta = 0$ and $\tan\alpha = 0$, so the origin is a cusp of the second kind. This can be seen from the explicit form of the equation $y = x^2(1 \pm \sqrt{x})$. y is not defined for $x < 0$, while for $0 < x < 1$ both values of y are positive; at the origin the tangent is horizontal.

4. Algebraic Curves of Type $F(x,y) = 0$, $F(x,y)$ Polynomial in x and y

If the equation does not contain constant and first-degree terms, then the origin is a double point. The corresponding tangents can be determined by making the sum of the second-degree terms equal.

■ For the lemniscate (**Fig. 2.66**, p. 101) at the origin $(0,0)$ the equations of the tangent $y = \pm x$ follow from $x^2 - y^2 = 0$.

If the equation does not contain second-degree terms as well but contains cubic terms, then the origin is a triple point.

3.6.1.4 Asymptotes of Curves

1. Definition

An *asymptote* is a straight line to which the curve approaches while it moves away from the origin **(Fig. 3.232)**.

The curve can approach the line from one side **(Fig. 3.232a)**, or it can intersect it again and again **(Fig. 3.232b)**.


a) b)

Figure 3.232

Not every curve which goes infinitely far from the origin (infinite branch of the curve) has an asymptote. For instance the entire part of an improper rational expression is called an asymptotic approximation (see 1.1.7.2, p. 15).

2. Functions Given in Parametric Form $x = x(t), y = y(t)$

To determine the equation of the asymptote first those values are determined for which $t \to t_i$ yields either $x(t) \to \pm\infty$ or $y(t) \to \pm\infty$ (or both).

There are the following cases:

a) $x(t) \to \infty$ for $t \to t_i$ but $y(t_i) = a \neq \infty$: $y = a$. The asymptote is a horizontal line. (3.515a)

b) $y(t) \to \infty$ for $t \to t_i$ but $x(t_i) = a \neq \infty$: $x = a$. The asymptote is a vertical line. (3.515b)

c) If both $y(t)$ and $x(t)$ tend to $\pm\infty$, then the following limits are to be calculated: $k = \lim\limits_{t \to t_i} \frac{y(t)}{x(t)}$ and $b = \lim\limits_{t \to t_i} [y(t) - kx(t)]$. If both exist, the equation of the asymptote is

$$y = kx + b. \tag{3.515c}$$

■ $x = \frac{m}{\cos t}$, $y = n(\tan t - t)$, $t_1 = \frac{\pi}{2}$, $t_2 = -\frac{\pi}{2}$, etc. Determine the asymptote at t_1:
$x(t_1) = y(t_1) = \infty$, $k = \lim\limits_{t \to \pi/2} \frac{n}{m}(\sin t - t\cos t) = \frac{n}{m}$,

$b = \lim\limits_{t \to \pi/2} \left[n(\tan t - t) - \frac{n}{m}\frac{m}{\cos t}\right] = n \lim\limits_{t \to \pi/2} \frac{\sin t - t\cos t - 1}{\cos t} = -\frac{n\pi}{2}$. For the asymptote (3.515c) gives $y = \frac{n}{m}x - \frac{n\pi}{2}$. For the second asymptote, etc. one gets similarly $y = -\frac{n}{m}x + \frac{n\pi}{2}$.

3. Functions Given in Explicit Form $y = f(x)$

The vertical asymptotes are at the points of discontinuity where the function $f(x)$ has an infinite jump (see 2.1.5.3, p. 59); the horizontal and oblique asymptotes have the equation

$$y = kx + b \text{ with } k = \lim_{x\to\infty} \frac{f(x)}{x}, \quad b = \lim_{x\to\infty}[f(x) - kx]. \tag{3.516}$$

4. Functions Given in Implicit Polynomial Form $F(x, y) = 0$

1. To determine the horizontal and vertical asymptotes one chooses the highest-degree terms with degree m from the polynomial expression in x and y, then one separates them as a function $\Phi(x, y)$ and solves, if it is possible, the equation $\Phi(x, y) = 0$ for x and y:

$$\Phi(x, y) = 0 \quad \text{yields} \quad x = \varphi(y), \quad y = \psi(x). \tag{3.517}$$

The values $y_1 = a$ for $x \to \infty$ give the horizontal asymptotes $y = a$; the values $x_1 = b$ for $y \to \infty$ the vertical ones $x = b$, if the limits exist.

2. To determine the oblique asymptotes the equation of the line $y = kx + b$ is to be substituted into the equation $F(x, y)$, then the resulting polynomial is to be ordered according to the powers of x:

$$F(x, kx + b) \equiv f_1(k, b)x^m + f_2(k, b)x^{m-1} + \cdots . \tag{3.518}$$

The parameters k and b can be yield, if they exist, from the equations

$$f_1(k, b) = 0, \quad f_2(k, b) = 0. \tag{3.519}$$

■ $x^3 + y^3 - 3axy = 0$ (Cartesian folium **Fig. 2.59**, 2.11.3, p. 96). Based on equation $F(x, kx + b) \equiv (1 + k^3)x^3 + 3(k^2b - ka)x^2 + \cdots$ according to (3.519), $f_1(k, b) = 1 + k^3 = 0$ and $f_2(k, b) = k^2b - ka = 0$ with the solutions $k = -1$, $b = -a$ the equation of the asymptote is $y = -x - a$.

3.6.1.5 General Discussion of a Curve Given by an Equation

Curves given by their equations (3.483)–(3.486) are investigated in order to know their properties and shapes.

1. Construction of a Curve Given by an Explicit Function $y = f(x)$

a) Determination of the domain (see 2.1.1, p. 48).

b) Determination of the symmetry of the curve with respect to the origin or to the y-axis checking if the function is odd or even (see 2.1.3.4, p. 51).

c) Determination of the behavior of the function at $\pm\infty$ by calculating the limits $\lim\limits_{x\to-\infty} f(x)$ and $\lim\limits_{x\to+\infty} f(x)$ (see 2.1.4.7, p. 55).

d) Determination of the points of discontinuity (see 2.1.5.3, p. 59).

e) Determination of the intersection points with the y-axis and with the x-axis calculating $f(0)$ and solving the equation $f(x) = 0$.

f) Determination of maxima and minima and finding the intervals of monotonicity where the function is increasing or decreasing.

g) Determination of inflection points and the equations of tangents at these points (see 3.6.1.3, p. 249).

With these data the graph of the function can be sketched, and where it is necessary, some values can be calculated to make it more precise.

■ Sketch the graph of the function $y = \dfrac{2x^2 + 3x - 4}{x^2}$:

a) The function is defined for all x except $x = 0$.

b) There is no symmetry.

c) For $x \to -\infty$ follows $y \to 2$, and obviously $y = 2 - 0$, i.e., approach from below, while $x \to \infty$ also holds $y \to 2$, but $y = 2 + 0$, an approach from above.

d) $x = 0$ is a point of discontinuity such that the function from left and also from right tends to $-\infty$, because y is negative for small values of x.

e) Because $f(0) = -\infty$ holds, there is no intersection point with the y-axis, and from $f(x) = 2x^2 + 3x - 4 = 0$ the intersection points with the x-axis are at $x_1 \approx 0.85$ and $x_2 \approx -2.35$.

f) A maximum is at $x = 8/3 \approx 2.66$ and here $y \approx 2.56$.

g) An inflection point is at $x = 4$, $y = 2.5$ with the slope of the tangent line $\tan\alpha = -1/16$.

h) After sketching the graph of the function based on these data **(Fig. 3.233)** one can calculate the intersection point of the curve and the asymptote, which is at $x = 4/3 \approx 1.33$ and $y = 2$.

2. Construction of a Curve Given by an Implicit Function $F(x, y) = 0$

There are no general rules for this case, because depending of the actual form of the function different steps can be or cannot be performed. If it is possible, the following steps are recommended:

a) Determination of all the intersection points with the coordinate axes.

b) Determination of the symmetry of the curve, replacing x by $-x$ and y by $-y$.

c) Determination of maxima and minima with respect to the x-axis and then interchanging x and y also with respect to the y-axis (see 6.1.5.3, p. 444).

d) Determination of the inflection points and the slopes of tangents there (see 3.6.1.3, p. 249).

e) Determination of singular points (see 3.6.1.3, **3.**, p. 250).

f) Determination of vertices (see 3.6.1.3, **2.**, p. 250) and the corresponding circles of curvature (see 3.6.1.2, **4.**, p. 246). It often happens that the curve's arc can hardly be distinguish from the circular segment of the circle of curvature on a relatively large segment.

g) Determination of the equations of asymptotes (see 3.6.1.4, p. 252) and the position of the curve branches related to the asymptotes.

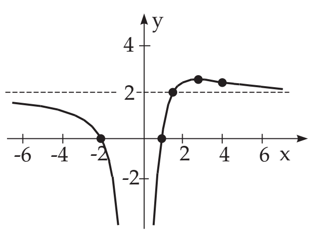


Figure 3.233

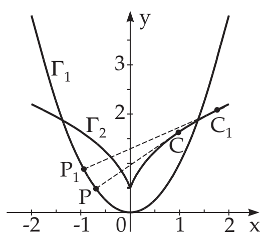


Figure 3.234

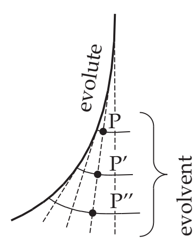


Figure 3.235

3.6.1.6 Evolutes and Evolvents

1. Evolute

The evolute is a second curve which is the locus of the centers of circles of curvature of the first curve (see 3.6.1.3, **5.**, p. 248); at the same time it is the envelope of the normals of the first curve (see also

3.6.1.7, p. 255). The parametric form of the evolute can be derived from (3.503) or (3.504) for the center of curvature if x_C and y_C are considered as running coordinates. If it is possible to eliminate the parameter (x, t or φ) from (3.502), (3.503), (3.504), then the equation of the evolute is obtained in Cartesian coordinates.

■ Determine the evolute of the parabola $y = x^2$ **(Fig. 3.234)**. From $X = x - \frac{2x(1+4x^2)}{2} = -4x^3$, $Y = x^2 + \frac{1+4x^2}{2} = \frac{1+6x^2}{2}$ and considering X and Y as running coordinates the evolute $Y = \frac{1}{2} + 3\left(\frac{X}{4}\right)^{2/3}$.

2. Evolvent or Involute

The evolvent, also called involute, of a curve Γ_2 is a curve Γ_1, whose evolute is Γ_2. Here every normal PC of the evolvent is a tangent of the evolute **(Fig. 3.234)**, and the length of arc $\overset{\frown}{CC_1}$ of the evolute is equal to the increment of the radius of curvature of the evolvent:

$$\overset{\frown}{CC_1} = \overline{P_1C_1} - \overline{PC}. \tag{3.520}$$

These properties show that the evolvent Γ_1 can be regarded as the curve traced by the end of a stretched thread unspooling from Γ_2. A given evolute corresponds to a family of curves, where every curve is determined by the initial length of the thread **(Fig. 3.235)**.
The equation of the evolute is obtained by the integration of a system of differential equations corresponding to its evolute. For the equation of the evolvent of the circle see 2.14.4, p. 106.

■ The catenoid is the evolute of the tractrix; the tractrix is the evolvent of the catenoid (see 2.15.1, p. 108).

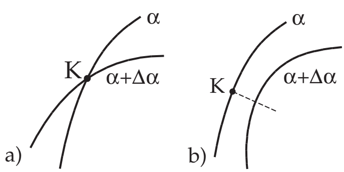


Figure 3.236

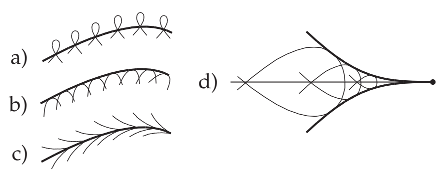


Figure 3.237

3.6.1.7 Envelope of a Family of Curves

1. Characteristic Point

Consider the one-parameter family of curves with equation

$$F(x, y, \alpha) = 0. \tag{3.521}$$

Every two close curves of this family corresponding to the values of parameter α and $\alpha+\Delta\alpha$ have *points K of nearest approach*. Such a point is either a point of intersection of the curves (α) and ($\alpha + \Delta\alpha$) or a point of the curve (α) whose distance from the curve ($\alpha + \Delta\alpha$) along the normal is an infinitesimal quantity of higher order than $\Delta\alpha$ **(Fig. 3.236a,b)**. For $\Delta\alpha \to 0$ the curve ($\alpha+\Delta\alpha$) tends to the curve (α), where in some cases the point K approaches a limit position, the *characteristic point*.

2. Geometric Locus of the Characteristic Points of a Family of Curves

Equation (3.521) may represent one or more curves. They are formed by the approaching points or by the boundary points of the family **(Fig. 3.237a)**, or they form an *envelope* of the family, i.e., a curve which contacts tangentially every curve of the family **(Fig. 3.237b)**. A combination of these two cases is also possible **(Fig. 3.237c,d)**.

3. Equation of the Envelope

The equation of the envelope can be calculated from (3.521), where α can be eliminated from the following equation system:

$$F = 0, \qquad \frac{\partial F}{\partial \alpha} = 0. \tag{3.522}$$

■ Determine the equation of the family of straight lines arising when the ends of a line segment AB with $|AB| = l$ are sliding along the coordinate axes **(Fig. 3.238a)**. The equation of the family of curves is $\dfrac{x}{l\sin\alpha} + \dfrac{y}{l\cos\alpha} = 1$ or $F \equiv x\cos\alpha + y\sin\alpha - l\sin\alpha\cos\alpha = 0$, $\dfrac{\partial F}{\partial\alpha} = -x\sin\alpha + y\cos\alpha - l\cos^2\alpha + l\sin^2\alpha = 0$. Eliminating α gives $x^{2/3} + y^{2/3} = l^{2/3}$ as an envelope, which is an astroid (**Fig. 3.238b**, see also p. 103).

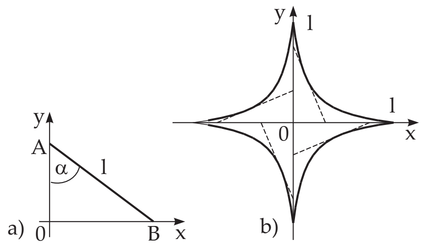


Figure 3.238

3.6.2 Space Curves

3.6.2.1 Ways to Define a Space Curve

1. Coordinate Equations

To define a space curve there are the following possibilities:

a) **Intersection of Two Surfaces:** $F(x, y, z) = 0, \quad \Phi(x, y, z) = 0.$ (3.523)

b) **Parametric Form:** $x = x(t), \quad y = y(t), \quad z = z(t),$ (3.524)

with t as an arbitrary parameter; mostly $t = x$, y or z are used.

c) **Parametric Form:** $x = x(s), \quad y = y(s), \quad z = z(s),$ (3.525a)

with the arc length s between a fixed point A and the running point P as the parameter:

$$s = \int_{t_0}^{t} \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2 + \left(\frac{dz}{dt}\right)^2}\, dt. \tag{3.525b}$$

2. Vector Equations

With $\vec{\mathbf{r}}$ as radius vector of an arbitrary point of the curve (see 3.5.1.1, **6.**, p. 182) the equation (3.524) can be written in the form

$$\vec{\mathbf{r}} = \vec{\mathbf{r}}(t), \quad \text{where} \quad \vec{\mathbf{r}}(t) = x(t)\vec{\mathbf{i}} + y(t)\vec{\mathbf{j}} + z(t)\vec{\mathbf{k}}, \tag{3.526}$$

and (3.525a) in the form

$$\vec{\mathbf{r}} = \vec{\mathbf{r}}(s), \quad \text{where} \quad \vec{\mathbf{r}}(s) = x(s)\vec{\mathbf{i}} + y(s)\vec{\mathbf{j}} + z(s)\vec{\mathbf{k}}. \tag{3.527}$$

3. Positive Direction

This is the direction of increasing parameter t for a curve given in the form (3.524) and (3.526); for (3.525a) and (3.527) it is the direction in which the arc length increases.

3.6.2.2 Moving Trihedral

1. Definitions

At every point P of a space curve can be defined three lines and three planes, apart from singular points. They intersect each other at the point P, and they are perpendicular to each other **(Fig.3.239).**

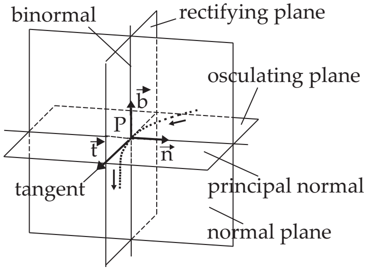


Figure 3.239

1. Tangent is the limiting position of the secants PN for $N \to P$ **(Fig.3.220)**, p. 244.

2. Normal Plane is a plane perpendicular to the tangent **(3.239)**. Every line passing through the point P and contained by this plane is called a normal of the curve at the point P.

3. Osculating Plane is the limiting position of the planes passing through three neighboring points M, P and N, for $N \to P$ and $M \to P$ **(Fig.3.240)**. The tangent line is contained by the osculating plane.

4. Principal Normal is the intersection line of the normal and the osculating plane, i.e., it is the normal contained by the osculating plane.

5. Binormal is the line perpendicular to the osculating plane through the point P.

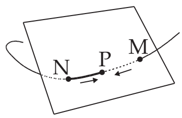


Figure 3.240

6. Rectifying Plane is the plane spanned by the tangent and binormal lines.

The positive directions on the lines (1.), (4.) and (5.) are defined as follows:

7. Moving trihedral The positive directions on the lines tangent, principal normal and binormal are defined as follows:

a) On the tangent line it is given by the positive direction of the curve; the unit tangent vector $\vec{\mathbf{t}}$ has this direction.

b) On the principal normal it is given by the sign of the curvature of the curve, and given by the unit normal vector $\vec{\mathbf{n}}$.

c) On the binormal it is defined by the unit vector

$$\vec{\mathbf{b}} = \vec{\mathbf{t}} \times \vec{\mathbf{n}}, \qquad (3.528)$$

where the three vectors $\vec{\mathbf{t}}$, $\vec{\mathbf{n}}$, and $\vec{\mathbf{b}}$ form a right-handed rectangular coordinate system, which is called the *moving trihedral*.

2. Position of the Curve Related to the Moving Trihedral

For the usual points of the curve the space curve is on one side of the rectifying plane at the point P, and intersects both the normal and osculating planes **(Fig.3.241a)**. The projections of a small segment of the curve at the point P on the three planes have approximately the following shapes:

1. On the osculating plane it is similar to a quadratic parabola **(Fig.3.241b)**.

2. On the rectifying plane it is similar to a cubic parabola **(Fig.3.241c)**.

3. On the normal plane it is similar to a semi cubical parabola **(Fig.3.241d)**.

If the curvature or the torsion of the curve are equal to zero at P or if P is a singular point, i.e., if $x'(t) = y'(t) = z'(t) = 0$ hold, then the curve may have a different shape.

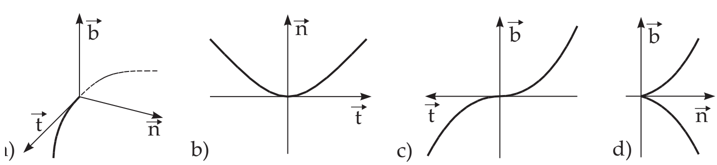


Figure 3.241

3. Equations of the Elements of the Moving Trihedral

1. The Curve is Defined in the Form (3.523) For the tangent see (3.529), for the normal plane see (3.530):

$$\frac{X-x}{\begin{vmatrix} \frac{\partial F}{\partial y} & \frac{\partial F}{\partial z} \\ \frac{\partial \Phi}{\partial y} & \frac{\partial \Phi}{\partial z} \end{vmatrix}} = \frac{Y-y}{\begin{vmatrix} \frac{\partial F}{\partial z} & \frac{\partial F}{\partial x} \\ \frac{\partial \Phi}{\partial z} & \frac{\partial \Phi}{\partial x} \end{vmatrix}} = \frac{Z-z}{\begin{vmatrix} \frac{\partial F}{\partial x} & \frac{\partial F}{\partial y} \\ \frac{\partial \Phi}{\partial x} & \frac{\partial \Phi}{\partial y} \end{vmatrix}}. \quad (3.529)$$

$$\begin{vmatrix} X-x & Y-y & Z-z \\ \frac{\partial F}{\partial x} & \frac{\partial F}{\partial y} & \frac{\partial F}{\partial z} \\ \frac{\partial \Phi}{\partial x} & \frac{\partial \Phi}{\partial y} & \frac{\partial \Phi}{\partial z} \end{vmatrix} = 0. \quad (3.530)$$

Here x, y, z are the coordinates of the point P of the curve and X, Y, Z are the running coordinates of the tangent or the normal plane; the partial derivatives belong to the point P.

2. The Curve is Defined in the Form (3.524), (3.526) In **Table 3.27** the coordinate and vector equations belonging to the point P are given with x, y, z and also with $\vec{\mathbf{r}}$. The running coordinates and the radius vector of the running point are denoted by X, Y, Z and $\vec{\mathbf{R}}$. The derivatives with respect to the parameter t refer to the point P.

3. The Curve is Defined in the Form (3.525a, 3.527) If the parameter is the arc length s, for the tangent and binormal, and for the normal and osculating plane the same equations are valid as in case 2, just t must be replaced by s. The equations of the principal normal and the rectifying plane will be simpler **(Table 3.28)**.

3.6.2.3 Curvature and Torsion

1. Curvature of a Curve and Radius of Curvature

The *curvature of a curve* at the point P is a measure which describes the deviation of the curve from a straight line in the very close neighborhood of this point. The exact definition is by the help of the tangent vector $\vec{\mathbf{t}} = \frac{d\vec{\mathbf{r}}}{ds}$ (see **Fig.3.242**):

$$K = \lim_{\overset{\frown}{PN} \to 0} \frac{\Delta \vec{\mathbf{t}}}{\overset{\frown}{PN}} = \frac{d\vec{\mathbf{t}}}{ds}. \quad (3.531)$$

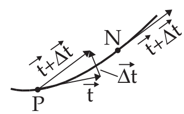


Figure 3.242

1. Radius of Curvature The radius of curvature is the reciprocal value of the absolute value of the curvature:

$$\rho = \frac{1}{|K|}. \quad (3.532)$$

2. Formulas to calculate K and ρ

a) If the curve is defined in the form (3.525a):

$$|K| = \left| \frac{d^2\vec{\mathbf{r}}}{ds^2} \right| = \sqrt{x''^2 + y''^2 + z''^2}, \quad (3.533)$$

where the derivatives are with respect to s.

b) If the curve is defined in the form (3.524):

$$K^2 = \frac{\left(\frac{d\vec{\mathbf{r}}}{dt}\right)^2 \left(\frac{d^2\vec{\mathbf{r}}}{dt^2}\right)^2 - \left(\frac{d\vec{\mathbf{r}}}{dt}\frac{d^2\vec{\mathbf{r}}}{dt^2}\right)^2}{\left|\left(\frac{d\vec{\mathbf{r}}}{dt}\right)^2\right|^3} = \frac{(x'^2 + y'^2 + z'^2)(x''^2 + y''^2 + z''^2) - (x'x'' + y'y'' + z'z'')^2}{(x'^2 + y'^2 + z'^2)^3}. \quad (3.534)$$

The derivatives are calculated here with respect to t.

Table 3.27 Vector and coordinate equations of accompanying configurations of a space curve

Vector equation	Coordinate equation
Tangent:	
$\vec{\mathbf{R}} = \vec{\mathbf{r}} + \lambda \frac{d\vec{\mathbf{r}}}{dt}$	$\frac{X-x}{x'} = \frac{Y-y}{y'} = \frac{Z-z}{z'}$
Normal plane:	
$(\vec{\mathbf{R}} - \vec{\mathbf{r}}) \frac{d\vec{\mathbf{r}}}{dt} = 0$	$x'(X-x) + y'(Y-y) + z'(Z-z) = 0$
Osculating plane:	
$(\vec{\mathbf{R}} - \vec{\mathbf{r}}) \frac{d\vec{\mathbf{r}}}{dt} \frac{d^2\vec{\mathbf{r}}}{dt^2} = 0$ [1]	$\begin{vmatrix} X-x & Y-y & Z-z \\ x' & y' & z' \\ x'' & y'' & z'' \end{vmatrix} = 0$
Binormal:	
$\vec{\mathbf{R}} = \vec{\mathbf{r}} + \lambda \left(\frac{d\vec{\mathbf{r}}}{dt} \times \frac{d^2\vec{\mathbf{r}}}{dt^2} \right)$	$\frac{X-x}{\begin{vmatrix} y' & z' \\ y'' & z'' \end{vmatrix}} = \frac{Y-y}{\begin{vmatrix} z' & x' \\ z'' & x'' \end{vmatrix}} = \frac{Z-z}{\begin{vmatrix} x' & y' \\ x'' & y'' \end{vmatrix}}$
Rectifying plane:	
$(\vec{\mathbf{R}} - \vec{\mathbf{r}}) \frac{d\vec{\mathbf{r}}}{dt} \left(\frac{d\vec{\mathbf{r}}}{dt} \times \frac{d^2\vec{\mathbf{r}}}{dt^2} \right) = 0$ [1]	$\begin{vmatrix} X-x & Y-y & Z-z \\ x' & y' & z' \\ l & m & n \end{vmatrix} = 0,$ with $l = y'z'' - y''z'$, $m = z'x'' - z''x'$, $n = x'y'' - x''y'$
Principal normal:	
$\vec{\mathbf{R}} = \vec{\mathbf{r}} + \lambda \frac{d\vec{\mathbf{r}}}{dt} \times \left(\frac{d\vec{\mathbf{r}}}{dt} \times \frac{d^2\vec{\mathbf{r}}}{dt^2} \right)$	$\frac{X-x}{\begin{vmatrix} y' & z' \\ m & n \end{vmatrix}} = \frac{Y-y}{\begin{vmatrix} z' & x' \\ n & l \end{vmatrix}} = \frac{Z-z}{\begin{vmatrix} x' & y' \\ l & m \end{vmatrix}}$

$\vec{\mathbf{r}}$ position vector of the space curve, $\vec{\mathbf{R}}$ position vector of the accomp. configuration

[1] For the mixed product of three vectors see 3.5.1.6, p. 185

Tabelle 3.28 Vector and coordinate equations of accompanying configurations as functions of the arc length

Element of trihedral	Vector equation	Coordinate equation
Principal normal	$\vec{\mathbf{R}} = \vec{\mathbf{r}} + \lambda \frac{d^2\vec{\mathbf{r}}}{ds^2}$	$\frac{X-x}{x''} = \frac{Y-y}{y''} = \frac{Z-z}{z''}$
Rectifying plane	$(\vec{\mathbf{R}} - \vec{\mathbf{r}}) \frac{d^2\vec{\mathbf{r}}}{ds^2} = 0$	$x''(X-x) + y''(Y-y) + z''(Z-z) = 0$

$\vec{\mathbf{r}}$ position vector of the space curve, $\vec{\mathbf{R}}$ position vector of the accomp. configuration

3. Helix The equations

$$x = a\cos t, \quad y = a\sin t, \quad z = bt \quad (a > 0, b > 0) \tag{3.535}$$

describe a so-called *helix* **(Fig.3.243)** as a *right screw*. If the observer is looking into the positive direction of the z-axis, which is at the same time the axis of the screw, then the screw climbs in a counter-clockwise direction. A helix winding itself in the opposite orientation is called a *left screw*.

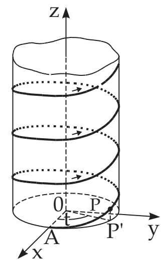


Figure 3.243

■ Determine the curvature of the helix (3.535). The parameter t is to be replaced by $s = t\sqrt{a^2+b^2}$. Then holds $x = a\cos\dfrac{s}{\sqrt{a^2+b^2}}$, $y = a\sin\dfrac{s}{\sqrt{a^2+b^2}}$, $z = \dfrac{bs}{\sqrt{a^2+b^2}}$, and according to (3.533), $K = \dfrac{a}{a^2+b^2}$, $\rho = \dfrac{a^2+b^2}{a}$. Both quantities K and ρ are constants.

Another method, without the parameter transformation in (3.534), produces the same result.

2. Torsion of a Curve and Radius of the Torsion Circe

The *torsion of a curve* at the point P is a measure which describes the deviation of the curve from a plane curve in the very close neighborhood of this point. The exact definition by the help of the binormal vector $\vec{\mathbf{b}}$ (3.528), p. 257 (see also **Fig.3.244**) is:

$$T = \lim_{\overset{\frown}{PN}\to 0} \frac{\Delta\vec{\mathbf{b}}}{\overset{\frown}{PN}} = \frac{d\vec{\mathbf{b}}}{ds}. \tag{3.536}$$

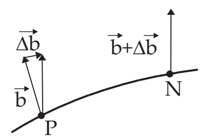


Figure 3.244

The *radius of torsion* is

$$\tau = \frac{1}{|T|}. \tag{3.537}$$

1. Formulas for Calculating T and τ

a) If the curve is defined in the form (3.525a):

$$T = \rho^2 \left(\frac{d\vec{\mathbf{r}}}{ds} \frac{d^2\vec{\mathbf{r}}}{ds^2} \frac{d^3\vec{\mathbf{r}}}{ds^3} \right)^* = \frac{\begin{vmatrix} x' & y' & z' \\ x'' & y'' & z'' \\ x''' & y''' & z''' \end{vmatrix}}{(x''^2 + y''^2 + z''^2)}, \tag{3.538}$$

where the derivatives are taken with respect to s.

b) If the curve is defined in the form (3.524):

$$T = \rho^2 \frac{\dfrac{d\vec{\mathbf{r}}}{dt} \dfrac{d^2\vec{\mathbf{r}}}{dt^2} \dfrac{d^3\vec{\mathbf{r}}}{dt^3}{}^*}{\left| \left(\dfrac{d\vec{\mathbf{r}}}{dt} \right)^2 \right|^3} = \rho^2 \frac{\begin{vmatrix} x' & y' & z' \\ x'' & y'' & z'' \\ x''' & y''' & z''' \end{vmatrix}}{(x'^2 + y'^2 + z'^2)^3}, \tag{3.539}$$

where ρ should be calculated by (3.532) and (3.533).

The torsion calculated by (3.538), (3.539) can be positive or negative. In the case $T > 0$ an observer standing on the principal normal parallel to the binormal sees that the curve has a right turn; in the case $T < 0$ it has a left turn.

■ The torsion of a helix is constant. With the notation R for the right screw and L for the left screw

---

*For the mixed product of three vectors see 3.5.1.6, **2.**, p. 185.

the torsion is

$$T_R = \left(\frac{a^2+b^2}{a}\right)^2 \frac{\begin{vmatrix} -a\sin t & a\cos t & b \\ -a\cos t & -a\sin t & 0 \\ a\sin t & -a\cos t & 0 \end{vmatrix}}{[(-a\sin t)^2 + (a\cos t)^2 + b^2]^3} = \frac{b}{a^2+b^2}, \qquad \tau = \frac{a^2+b^2}{b}; \quad T_L = -\frac{b}{a^2+b^2}.$$

3. Frenet Formulas

The derivatives of the vectors $\vec{\mathbf{t}}$, $\vec{\mathbf{n}}$, and $\vec{\mathbf{b}}$ can be expressed by the Frenet formulas:

$$\frac{d\vec{\mathbf{t}}}{ds} = \frac{\vec{\mathbf{n}}}{\rho}, \quad \frac{d\vec{\mathbf{n}}}{ds} = -\frac{\vec{\mathbf{t}}}{\rho} + \frac{\vec{\mathbf{b}}}{\tau}, \quad \frac{d\vec{\mathbf{b}}}{ds} = -\frac{\vec{\mathbf{n}}}{\tau}. \tag{3.540}$$

Here ρ is the radius of curvature, and τ is the radius of torsion.

4. Darboux Vector

The Frenet formulas (3.540) can be represented in the clearly arranged form (Darboux formulas)

$$\frac{d\vec{\mathbf{t}}}{ds} = \vec{\mathbf{d}} \times \vec{\mathbf{t}}, \quad \frac{d\vec{\mathbf{n}}}{ds} = \vec{\mathbf{d}} \times \vec{\mathbf{n}}, \quad \frac{d\vec{\mathbf{b}}}{ds} = \vec{\mathbf{d}} \times \vec{\mathbf{b}}. \tag{3.541}$$

Here $\vec{\mathbf{d}}$ is the Darboux vector, which has the form

$$\vec{\mathbf{d}} = \frac{1}{\tau}\vec{\mathbf{t}} + \frac{1}{\rho}\vec{\mathbf{b}}. \tag{3.542}$$

Remarks:
1. With the help of the Darboux vector the Frenet formulas can be interpreted in the sense of kinematics (see [3.4]).
2. The modulus of the Darboux vector equals the so-called *total curvator* λ of a space curve:

$$\lambda = \sqrt{\frac{1}{\rho^2} + \frac{1}{\tau^2}} = |\vec{\mathbf{d}}|. \tag{3.543}$$

3.6.3 Surfaces

3.6.3.1 Ways to Define a Surface

1. Equation of a Surface

Surfaces can be defined in different ways:

a) **Implicit Form** : $F(x,y,z) = 0.$ (3.544)

b) **Explicit Form** : $z = f(x,y).$ (3.545)

c) **Parametric Form** : $x = x(u,v), \quad y = y(u,v), \quad z = z(u,v).$ (3.546)

d) **Vector Form** : $\vec{\mathbf{r}} = \vec{\mathbf{r}}(u,v)$ with $\vec{\mathbf{r}} = x(u,v)\vec{\mathbf{i}} + y(u,v)\vec{\mathbf{j}} + z(u,v)\vec{\mathbf{k}}.$ (3.547)

Running the parameters u and v over all allowed values gives the coordinates and the radius vectors of all points of the surface from (3.546) and (3.547). The elimination of the parameters u and v from the parametric form (3.546) (if possible) yields the implicit form (3.544). The explicit form (3.545) is a special case of the parametric form with $u = x$ and $v = y$.

■ The equation of the sphere in Cartesian coordinates, parametric form, and vector form **(Fig. 3.246)**:

$x^2 + y^2 + z^2 - a^2 = 0;$ (3.548a) $\qquad x = a\cos u \sin v,\ y = a\sin u \sin v,\ z = a\cos v;$ (3.548b)

$$\vec{\mathbf{r}} = a(\cos u \sin v\vec{\mathbf{i}} + \sin u \sin v\vec{\mathbf{j}} + \cos v\vec{\mathbf{k}}). \tag{3.548c}$$

2. Curvilinear Coordinates on a Surface

If a surface is given in the form (3.546) or (3.547), and the values of the parameter u can be changed while the other parameter $v = v_0$ is fixed, then the points $\vec{\mathbf{r}}(x,y,z)$ describe a curve $\vec{\mathbf{r}} = \vec{\mathbf{r}}(u, v_0)$ on

the surface. Substituting for v different but fixed values $v = v_1, v = v_2, \dots, v = v_n$ one after the other, gives a family of curves on the surface. When moving along a curve with $v =$ const only u is changing, this curve is called the *u-line* **(Fig. 3.245)**. Analogously one gets another family of curves, the *v-lines*, by varying v and keeping $u =$ const fixed with $u_1, u_2, \dots, u_n$. In this way a net of coordinate lines is defined on the surface (3.546), where the two fixed numbers $u = u_i$ and $v = v_k$ are the *curvilinear* or *Gauss coordinates* of the point P on the surface.
If a surface is given in the form (3.545), the coordinate lines are the intersection curves of the surface with the planes $x =$ const and $y =$ const. With equations in implicit form $F(u, v) = 0$ or with the parametric equations $u = u(t)$ and $v = v(t)$ of these coordinates, curves on the surfaces can be defined.

■ In the parametric equations of the sphere (3.548b,c) u means the *geographical longitude* of a point P, and v means its *polar distance*. The v lines are here the *meridians* APB; the u lines are the *parallel circles* CPD **(Fig. 3.246)**.

3.6.3.2 Tangent Plane and Surface Normal

1. Definitions

1. Tangent Plane The precise general mathematical definition of the tangent plane is rather complicated, so here the investigation is to be restricted to the case, when the surface is defined by two parameters. Suppose, for a neighborhood of the point $P(x, y, z)$, the mapping $(u, v) \to \vec{\mathbf{r}}(u, v)$ is invertible, the partial derivatives $\vec{\mathbf{r}}_u = \dfrac{\partial \vec{\mathbf{r}}}{\partial u}$ and $\vec{\mathbf{r}}_v = \dfrac{\partial \vec{\mathbf{r}}}{\partial v}$ are continuous, and not parallel to each other. Then $P(x, y, z)$ is called a regular point of the surface. If P is regular, then the tangents of all curves passing through P, and having a tangent here, are in the same plane, and this plane is called the *tangent plane* of the surface at P. If this happens, the partial derivatives $\vec{\mathbf{r}}_u$, $\vec{\mathbf{r}}_v$ are parallel (or zero) only for certain parametrization of the surface. If they are parallel for every parametrization, the point is called a *singular point* (see **3.**, p. 263).

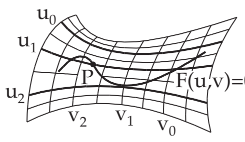


Figure 3.245

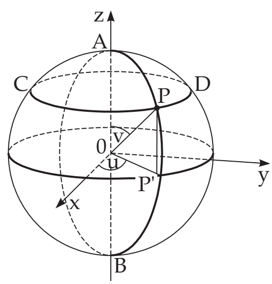


Figure 3.246

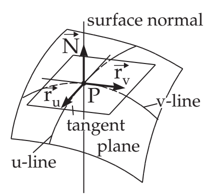


Figure 3.247

2. Surface Normal The line perpendicular to the tangent plane going through the point P is called *surface normal* at the point P **(Fig. 3.247)**.

3. Normal Vector The tangent plane is spanned by two vectors, by the tangent vectors

$$\vec{\mathbf{r}}_u = \frac{\partial \vec{\mathbf{r}}}{\partial u}, \quad \vec{\mathbf{r}}_v = \frac{\partial \vec{\mathbf{r}}}{\partial v} \tag{3.549a}$$

of the u- and v-lines. The vector product of the tangent vectors $\vec{\mathbf{r}}_u \times \vec{\mathbf{r}}_v$ is a vector in the direction of the surface normal. Its unit vector

$$\vec{\mathbf{N}}_0 = \frac{\vec{\mathbf{r}}_u \times \vec{\mathbf{r}}_v}{|\,\vec{\mathbf{r}}_u \times \vec{\mathbf{r}}_v\,|} \tag{3.549b}$$

is called the normal vector. Its direction to one or other side of the surface depends on which variable is the first and which one is the second coordinate among u and v.

2. Equations of the Tangent Plane and the Surface Normal (see **Table 3.29**)

■ **A:** For the sphere with equation (3.548a) is valid:

a) as tangent plane: $2x(X-x)+2y(Y-y)+2z(Z-z)=0$ or $xX+yY+zZ-a^2=0$, (3.550a)

b) as surface normal: $\dfrac{X-x}{2x}=\dfrac{Y-y}{2y}=\dfrac{Z-z}{2z}$ or $\dfrac{X}{x}=\dfrac{Y}{y}=\dfrac{Z}{z}$. (3.550b)

■ **B:** For the sphere with equation (3.548b) is valid:

a) as tangent plane: $X\cos u\sin v+Y\sin u\sin v+Z\cos v=a$, (3.550c)

b) as surface normal: $\dfrac{X}{\cos u\sin v}=\dfrac{Y}{\sin u\sin v}=\dfrac{Z}{\cos v}$. (3.550d)

3. Singular Points of the Surface

If for a point with coordinates $x=x_1$, $y=y_1$, $z=z_1$ of the surface with equation (3.544) all the equalities

$$\frac{\partial F}{\partial x}=0,\ \frac{\partial F}{\partial y}=0,\ \frac{\partial F}{\partial z}=0, F(x,y,z)=0 \tag{3.551}$$

are fulfilled, i.e., if every first-order derivative is zero at the point $P(x_1,y_1,z_1)$, then this point is called a *singular point*. All tangents going through here do not form a plane, but a cone of second order with the equation

$$\frac{\partial^2F}{\partial x^2}(X-x_1)^2+\frac{\partial^2F}{\partial y^2}(Y-y_1)^2+\frac{\partial^2F}{\partial z^2}(Z-z_1)^2+2\frac{\partial^2F}{\partial x\partial y}(X-x_1)(Y-y_1)$$
$$+2\frac{\partial^2F}{\partial y\partial z}(Y-y_1)(Z-z_1)+2\frac{\partial^2F}{\partial z\partial x}(Z-z_1)(X-x_1)=0, \tag{3.552}$$

where the derivatives belong to the point $P(x_1,y_1,z_1)$. If also the second derivatives are equal to zero at this point, then there is a more complicated type of singularity. Then the tangents form a cone of third or even higher order.

3.6.3.3 Line Elements of a Surface

1. Differential of Arc

Consider a surface given in the form (3.546) or (3.547). Let $P(u,v)$ an arbitrary point and $N(u+du,v+dv)$ another one close to P, both on the surface. The arclength of the arc segment $\overset{\frown}{PN}$ on the surface can be approximately calculated by the *differential of an arc* or the *line element of the surface* with the formula

$$ds^2=E\,du^2+2F\,du\,dv+G\,dv^2, \tag{3.553a}$$

where the three coefficients

$$E=\vec{\mathbf{r}}_u^2=\left(\frac{\partial x}{\partial u}\right)^2+\left(\frac{\partial y}{\partial u}\right)^2+\left(\frac{\partial z}{\partial u}\right)^2,\quad F=\vec{\mathbf{r}}_u\vec{\mathbf{r}}_v=\frac{\partial x}{\partial u}\frac{\partial x}{\partial v}+\frac{\partial y}{\partial u}\frac{\partial y}{\partial v}+\frac{\partial z}{\partial u}\frac{\partial z}{\partial v},$$
$$G=\vec{\mathbf{r}}_v^2=\left(\frac{\partial x}{\partial v}\right)^2+\left(\frac{\partial y}{\partial v}\right)^2+\left(\frac{\partial z}{\partial v}\right)^2 \tag{3.553b}$$

are calculated at the point P. The right-hand side (3.553a) is called the *first quadratic fundamental form of the surface.*

■ **A:** For the sphere given in the form (3.548c) clearly

$$E=a^2\sin^2v,\quad F=0,\quad G=a^2,\quad ds^2=a^2(sin^2v\,du^2+dv^2). \tag{3.554}$$

■ **B:** For a surface given in the form (3.545) clearly

$$E = 1 + p^2, \quad F = pq, \quad G = 1 + q^2 \quad \text{with} \quad p = \frac{\partial z}{\partial x}, \quad q = \frac{\partial z}{\partial y}. \tag{3.555}$$

Table 3.29 Equations of the tangent plane and surface normal

Type of equation	Tangent plane	Surface normal
(3.544)	$\frac{\partial F}{\partial x}(X-x) + \frac{\partial F}{\partial y}(Y-y) + \frac{\partial F}{\partial z}(Z-z) = 0$	$\dfrac{X-x}{\frac{\partial F}{\partial x}} = \dfrac{Y-y}{\frac{\partial F}{\partial y}} = \dfrac{Z-z}{\frac{\partial F}{\partial z}}$
(3.545)	$Z - z = p(X-x) + q(Y-y)$	$\dfrac{X-x}{p} = \dfrac{Y-y}{q} = \dfrac{Z-z}{-1}$
(3.546)	$\begin{vmatrix} X-x & Y-y & Z-z \\ \frac{\partial x}{\partial u} & \frac{\partial y}{\partial u} & \frac{\partial z}{\partial u} \\ \frac{\partial x}{\partial v} & \frac{\partial y}{\partial v} & \frac{\partial z}{\partial v} \end{vmatrix} = 0$	$\dfrac{X-x}{\begin{vmatrix} \frac{\partial y}{\partial u} & \frac{\partial z}{\partial u} \\ \frac{\partial y}{\partial v} & \frac{\partial z}{\partial v} \end{vmatrix}} = \dfrac{Y-y}{\begin{vmatrix} \frac{\partial z}{\partial u} & \frac{\partial x}{\partial u} \\ \frac{\partial z}{\partial v} & \frac{\partial x}{\partial v} \end{vmatrix}} = \dfrac{Z-z}{\begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial y}{\partial u} \\ \frac{\partial x}{\partial v} & \frac{\partial y}{\partial v} \end{vmatrix}}$
(3.547)	$(\vec{\mathbf{R}} - \vec{\mathbf{r}})\vec{\mathbf{r_1}}\vec{\mathbf{r_2}} = 0^{*1)}$ or $(\vec{\mathbf{R}} - \vec{\mathbf{r}})\vec{\mathbf{N}} = 0$	$\vec{\mathbf{R}} = \vec{\mathbf{r}} + \lambda(\vec{\mathbf{r_1}} \times \vec{\mathbf{r_2}})$ or $\vec{\mathbf{R}} = \vec{\mathbf{r}} + \lambda\vec{\mathbf{N}}$

In this table x, y, z and $\vec{\mathbf{r}}$ are the coordinates and radius vector of the points P of the curve; X, Y, Z and $\vec{\mathbf{R}}$ are the running coordinates and radius vectors of the points of the tangent plane and surface normal in the point P; furthermore $p = \frac{\partial z}{\partial x}$, $q = \frac{\partial z}{\partial y}$ and $\vec{\mathbf{N}}$ is the normal vector. Besides $\vec{\mathbf{r_1}} = \frac{\partial \vec{\mathbf{r}}}{\partial u}$ and $\vec{\mathbf{r_2}} = \frac{\partial \vec{\mathbf{r}}}{\partial v}$; $p = \frac{\partial z}{\partial x}$, $q = \frac{\partial z}{\partial y}$.

[1] For the mixed product of three vectors see 3.5.1.6, **2.**, p. 185

2. Measurement on the Surface

1. The Arclength of a surface curve $u = u(t)$, $v = v(t)$ for $t_0 \le t \le t_1$ is calculated by the formula

$$L = \int_{t_0}^{t_1} ds = \int_{t_0}^{t_1} \sqrt{E\left(\frac{du}{dt}\right)^2 + 2F\frac{du}{dt}\frac{dv}{dt} + G\left(\frac{dv}{dt}\right)^2}\, dt. \tag{3.556}$$

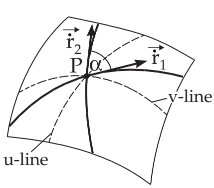


Figure 3.248

2. The Angle Between Two Curves $\vec{\mathbf{r}}_1 = \vec{\mathbf{r}}(u_1(t), v_1(t))$ and $\vec{\mathbf{r}}_2 = \vec{\mathbf{r}}(u_2(t), v_2(t))$ on the surface $\vec{\mathbf{r}} = \vec{\mathbf{r}}(u, v)$ is the angle between their tangents with the direction vectors $\dot{\vec{\mathbf{r}}}_1$ and $\dot{\vec{\mathbf{r}}}_2$ (**Fig. 3.248**). It is given by the formula

$$\cos\alpha = \frac{\dot{\vec{\mathbf{r}}}_1 \dot{\vec{\mathbf{r}}}_2}{\sqrt{\dot{\vec{\mathbf{r}}}_1^2 \dot{\vec{\mathbf{r}}}_2^2}} = \frac{E\dot{\vec{\mathbf{u}}}_1^2 \dot{\vec{\mathbf{u}}}_2^2 + F(\dot{\vec{\mathbf{u}}}_1\dot{\vec{\mathbf{v}}}_2 + \dot{\vec{\mathbf{v}}}_1\dot{\vec{\mathbf{u}}}_2) + G\dot{\vec{\mathbf{v}}}_1\dot{\vec{\mathbf{v}}}_2}{\sqrt{E\dot{\vec{\mathbf{u}}}_1^2 + 2F\dot{\vec{\mathbf{u}}}_1\dot{\vec{\mathbf{v}}}_1 + G\dot{\vec{\mathbf{v}}}_1^2}\sqrt{E\dot{\vec{\mathbf{u}}}_2^2 + 2F\dot{\vec{\mathbf{u}}}_2\dot{\vec{\mathbf{v}}}_2 + G\dot{\vec{\mathbf{v}}}_2^2}}. \tag{3.557}$$

Here the coefficients E, F and G are calculated at the point P and $\dot{\vec{\mathbf{u}}}_1$, $\dot{\vec{\mathbf{u}}}_2$, $\dot{\vec{\mathbf{v}}}_1$ and $\dot{\vec{\mathbf{v}}}_2$ represent the first derivatives of $u_1(t)$, $u_2(t)$, $v_1(t)$ and $v_2(t)$, calculated for the value of the parameter t at the point P.

If the numerator of (3.557) vanishes, the two curves are perpendicular to each other. The condition of orthogonality for the coordinate lines $v =$ const and $u =$ const is $F = 0$.

3. The Area of a Surface Patch S bounded by an arbitrary curve which is on the surface can be calculated by the double integral

$$S = \int\limits_{(S)} dS \qquad (3.558a) \qquad \text{with} \quad dS = \sqrt{EG - F^2}\,du\,dv. \qquad (3.558b)$$

dS is called the *surface element* or *element of area*.

The calculation of length, angle, and area on a surface is possible with (3.556), (3.557), (3.558a,b) if the coefficients E, F, and G of the first fundamental form are known. So the first quadratic fundamental form defines a *metric on the surface*.

3. Applicability of Surfaces by Bending

If a surface is deformed by bending, without stretching, compression or tearing, then its metric remains unchanged. In other words, the first quadratic fundamental form is invariant under bendings. Two surfaces having the same first quadratic fundamental form can be rolled onto each other.

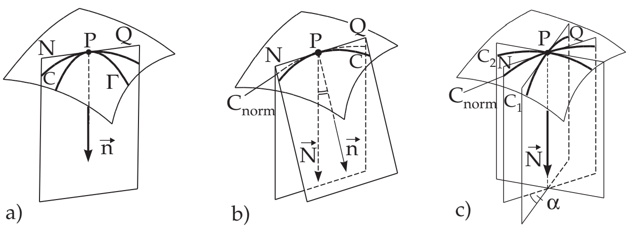


Figure 3.249

3.6.3.4 Curvature of a Surface

1. Curvatures of Curves on a Surface

If different curves Γ are drawn through a point P of the surface **(Fig. 3.249)**, their radii of curvature ρ at the point P are related as follows:

1. The Radius of Curvature ρ of a curve Γ at the point P is equal to the radius of curvature of a curve C, which is the intersection of the surface and the osculating plane of the curve Γ at the point P **(Fig. 3.249a)**.

2. Meusnier's Theorem For every plane section curve C of a surface **(Fig. 3.249b)** the radius of curvature can be calculated by the formula

$$\rho = R\cos(\vec{\mathbf{n}}, \vec{\mathbf{N}}). \qquad (3.559)$$

Here R is the radius of curvature of the *normal section* $C_{\rm norm}$, which goes through the same tangent NQ as C and also contains the unit vector $\vec{\mathbf{N}}$ of the surface normal; $\sphericalangle(n, N)$ is the angle between the unit vector $\vec{\mathbf{n}}$ of the *principal normal section* of the curve C and the unit vector $\vec{\mathbf{N}}$ of the surface normal. The sign of ρ in (3.559) is positive, if $\vec{\mathbf{N}}$ is on the concave side of the curve $C_{\rm norm}$ and negative otherwise.

3. Euler's Formula The curvature of a surface for every normal section $C_{\rm norm}$ at the point P can be calculated by the Euler formula

$$\frac{1}{R} = \frac{\cos^2\alpha}{R_1} + \frac{\sin^2\alpha}{R_2}, \qquad (3.560)$$

where R_1 and R_2 are the *radii of principal curvatures* (see (3.562a)) and α is the angle between the planes of the section C and C_1 **(Fig. 3.249c)**.

2. Radii of Principal Curvature

The *radii of principal curvatures* of a surface are the radii with maximum and minimum values. They can be calculated by the *principal normal sections* C_1 and C_2 **(Fig. 3.249c)**. The planes of C_1 and C_2 are perpendicular to each other, their directions are defined by the value $\frac{dy}{dx}$ which can be calculated from the quadratic equation

$$[tpq - s(1+q^2)]\left(\frac{dy}{dx}\right)^2 + [t(1+p^2) - r(1+q^2)]\frac{dy}{dx} + [s(1+p^2) - rpq] = 0, \tag{3.561}$$

where the parameters p, q, r, s, t are defined in (3.562b). If the surface is given in the explicit form (3.545), R_1 and R_2 are the roots of the quadratic equation

$$(rt - s^2)R^2 + h[2pqs - (1+p^2)t - (1+q^2)r]R + h^4 = 0 \quad \text{with} \tag{3.562a}$$

$$p = \frac{\partial z}{\partial x}, \quad q = \frac{\partial z}{\partial y}, \quad r = \frac{\partial^2 z}{\partial x^2}, \quad s = \frac{\partial^2 z}{\partial x \partial y}, \quad t = \frac{\partial^2 z}{\partial y^2} \quad \text{and} \quad h = \sqrt{1+p^2+q^2}. \tag{3.562b}$$

The signs of R, R_1 and R_2 can be determined by the same rule as in (3.559).
If the surface is given in vector form (3.547), then instead of (3.561) and (3.562a) the corresponding equations hold

$$(GM - FN)\left(\frac{dv}{du}\right)^2 + (GL - EN)\frac{dv}{du} + (FL - EM) = 0, \tag{3.563a}$$

$$(LN - M^2)R^2 - (EN - 2FM + GL)R + (EG - F^2) = 0, \tag{3.563b}$$

with the coefficients L, M, N of the *second quadratic fundamental form*. They are given by the equalities

$$L = \vec{\mathbf{r}}_{uu}\,\vec{\mathbf{R}} = \frac{d}{\sqrt{EG - F^2}}, \quad M = \vec{\mathbf{r}}_{uv}\,\vec{\mathbf{R}} = \frac{d'}{\sqrt{EG - F^2}}, \quad N = \vec{\mathbf{r}}_{vv}\,\vec{\mathbf{R}} = \frac{d''}{\sqrt{EG - F^2}}. \tag{3.563c}$$

Here the vectors $\vec{\mathbf{r}}_{uu}$, $\vec{\mathbf{r}}_{uv}$, and $\vec{\mathbf{r}}_{vv}$ are the second-order partial derivatives of the radius vector $\vec{\mathbf{r}}$ with respect to the parameters u and v. In the numerators there are the determinants

$$d = \begin{vmatrix} \frac{\partial^2 x}{\partial u^2} & \frac{\partial^2 y}{\partial u^2} & \frac{\partial^2 z}{\partial u^2} \\ \frac{\partial x}{\partial u} & \frac{\partial y}{\partial u} & \frac{\partial z}{\partial u} \\ \frac{\partial x}{\partial v} & \frac{\partial y}{\partial v} & \frac{\partial z}{\partial v} \end{vmatrix}, \quad d' = \begin{vmatrix} \frac{\partial^2 x}{\partial u \partial v} & \frac{\partial^2 y}{\partial u \partial v} & \frac{\partial^2 z}{\partial u \partial v} \\ \frac{\partial x}{\partial u} & \frac{\partial y}{\partial u} & \frac{\partial z}{\partial u} \\ \frac{\partial x}{\partial v} & \frac{\partial y}{\partial v} & \frac{\partial z}{\partial v} \end{vmatrix}, \quad d'' = \begin{vmatrix} \frac{\partial^2 x}{\partial v^2} & \frac{\partial^2 y}{\partial v^2} & \frac{\partial^2 z}{\partial v^2} \\ \frac{\partial x}{\partial u} & \frac{\partial y}{\partial u} & \frac{\partial z}{\partial u} \\ \frac{\partial x}{\partial v} & \frac{\partial y}{\partial v} & \frac{\partial z}{\partial v} \end{vmatrix}. \tag{3.563d}$$

The expression

$$L\,du^2 + 2M\,du\,dv + N\,dv^2 \tag{3.563e}$$

is called the second quadratic fundamental form. It contains the curvature properties of the surface.
Lines of curvature are the curves on the surface which have the direction of the principal normal section at every point. Their equations can be got by integrating (3.561) or (3.563a).

3. Classification of the Points of Surfaces

1. Elliptic and Umbilical Points If at a point P the radii R_1 and R_2 of the principal curvatures have the same sign, then every point of the surface is on the same side of the tangent plane in a close neighborhood of this point, and P is called an *elliptic point* **(Fig. 3.250a)** on the surface. This fact can be expressed analytically by the relation

$$LN - M^2 > 0. \tag{3.564a}$$

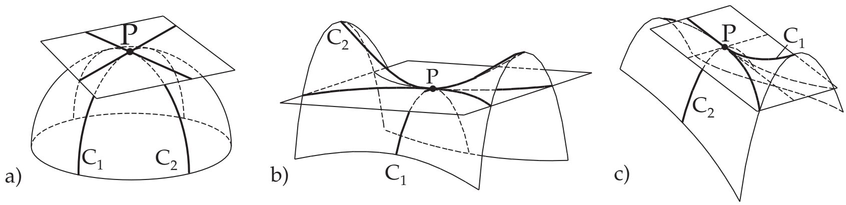


Figure 3.250

2. Circular or Umbilical Point This is the point P of the surface where the radii of the principal curvatures are equal:

$$R_1 = R_2. \qquad (3.564b)$$

Then for the normal sections $R = \text{const.}$

3. Hyperbolic Point In the case of different signs of the radii R_1 and R_2 of the principal curvatures the concave sides of the principal normal sections are in opposite directions. The tangent plane intersects the surface, so the surface has a saddle form in the neighborhood of P. P is called a *hyperbolic point* **(Fig. 3.250b)**; the analytic mark of this point is the relation

$$LN - M^2 < 0. \qquad (3.564c)$$

4. Parabolic Point If one of the two radii of principal curvatures R_1 or R_2 is equal to ∞, then either one of the principal normal sections has an inflection point here, or it is a straight line. At P there is a *parabolic point* **(Fig. 3.250c)** of the surface with the analytic mark

$$LN - M^2 = 0. \qquad (3.564d)$$

■ All the points of an ellipsoid are elliptic, those of a hyperboloid of one sheet are hyperbolic, and those of a cylinder are parabolic.

4. Curvature of a Surface

Two quantities are used mostly to characterize the curvature of a surface:

1. Mean Curvature of a surface at the point P: $H = \frac{1}{2}\left(\frac{1}{R_1} + \frac{1}{R_2}\right).$ (3.565a)

2. Gauss Curvature of a surface at the point P: $K = \frac{1}{R_1 R_2}.$ (3.565b)

■ **A:** For the circular cylinder with the radius a these are $H = \frac{1}{2a}$ and $K = 0$.

■ **B:** For elliptic points $K > 0$, for hyperbolic points $K < 0$, and for parabolic points $K = 0$ hold.

3. Calculation of *H* and *K*, if the surface is given in the form $z = f(x, y)$:

$$H = \frac{r(1+q^2) - 2pqs + t(1+p^2)}{2(1+p^2+q^2)^{3/2}}, \qquad (3.566a) \qquad K = \frac{rt - s^2}{(1+p^2+q^2)^2}. \qquad (3.566b)$$

For the meaning of p, q, r, s, t see (3.562b).

4. Classification of the Surfaces According to their Curvature

1. Minimal Surfaces are surfaces with a zero mean curvature H at every point, i.e., with $R_1 = -R_2$.

2. Surfaces of Constant Curvature have a constant Gauss curvature $K = \text{const.}$

■ **A:** $K > 0$, for instance the sphere.

■ **B:** $K < 0$, for instance the pseudosphere **(Fig. 3.251)**, i.e., the surface obtained by rotating the tractrix around the symmetry axis **(Fig. 2.79)**.

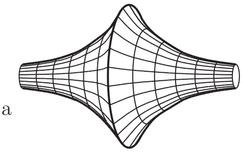


Figure 3.251

3.6.3.5 Ruled Surfaces and Developable Surfaces

1. Ruled Surface

A surface is called a *ruled surface* if it can be represented by moving a line in space.

2. Developable Surface

If a ruled surface can be developed upon a plane, i.e., rolled out without stretching or contracting any part of it, it is called a *developable surface*. Not every ruled surface is developable.
Developable surfaces have the following characteristic properties:

a) For all points the Gauss curvature is equal to zero, and

b) if the surface is given in the explicit form $z = f(x, y)$ the conditions of developability are fulfilled:

$$\text{a)}\quad K = 0, \qquad \text{b)}\quad rt - s^2 = 0. \tag{3.567}$$

For the meaning of r, t, and s see (3.562b).

■ **A:** The cone **(Fig. 3.192)** and cylinder **(Fig. 3.198)** are developable surfaces.

■ **B:** The hyperboloid of one sheet **(Fig. 3.196)** and hyperbolic paraboloid **(Fig. 3.197)** are ruled surfaces but they cannot be developed upon a plane.

3.6.3.6 Geodesic Lines on a Surface

1. Concept of Geodesic Line

(see also 3.4.1.1, **3.**, p. 160). A theoretic curve of the surface can go through every point $P(u, v)$ of the surface in every direction determined by the differential quotient $\dfrac{dv}{du}$, and it is called a *geodesic line*. It has the same role on the surface as the straight line on the plane, and it has the following properties:

1. Geodesic lines are the shortest curves between two points on the surface.

2. If a material point is moving on a surface drawn by another material point on the same surface, and no other forces have influence on it, then it is moving along a geodesic line.

3. If an elastic thread is stretched on a given surface, then it has the shape of a geodesic line.

2. Definition

A geodesic line on a surface is a curve such that its principal normal at every point has the same direction as the surface normal.

■ On a circular cylinder the geodesic lines are circular helices.

3. Equation of the Geodesic Line

If the surface is given in the explicit form $z = f(x, y)$, the differential equation of the geodesic lines is

$$(1 + p^2 + q^2)\frac{d^2y}{dx^2} = pt\left(\frac{dy}{dx}\right)^3 + (2ps - qt)\left(\frac{dy}{dx}\right)^2 + (pr - 2qs)\frac{dy}{dx} - qr. \tag{3.568}$$

If the surface is given in parametric form (3.546), then the differential equation of the geodesic lines is fairly complicated. For the meaning of p, q, r, s, and t see (3.562b).

4 Linear Algebra

4.1 Matrices

4.1.1 Notion of Matrix

1. Matrices A of Size (m, n) or Briefly $\mathbf{A}_{(m,n)}$

are systems of m times n elements, e.g., real or complex numbers, or functions, derivatives, vectors, arranged in m rows and n columns:

$$\mathbf{A} = (a_{ij}) = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix} \begin{matrix} \leftarrow & \text{1st row} \\ \leftarrow & \text{2nd row} \\ & \vdots \\ \leftarrow & m\text{-th row} \end{matrix} \tag{4.1}$$

$$\begin{matrix} \uparrow & \uparrow & & \uparrow \\ \text{1st} & \text{2nd} & & n\text{-th column.} \end{matrix}$$

With the notion *size of a matrix* matrices are classified according to their number of rows m and number of columns n: $\mathbf{A}$ of size (m, n). A matrix is called a *square matrix* if the number of rows and columns is equal, otherwise it is a *rectangular matrix*.

2. Real and Complex Matrices

Real matrices have real elements, *complex matrices* have complex elements. If a matrix has complex elements

$$a_{\mu\nu} + \mathrm{i}b_{\mu\nu} \tag{4.2a}$$

it can be decomposed into the form

$$\mathbf{A} + \mathrm{i}\,\mathbf{B} \tag{4.2b}$$

where $\mathbf{A}$ and $\mathbf{B}$ have real elements only (arithmetical operations see 4.1.4, p. 272).
If a matrix $\mathbf{A}$ has complex elements, then its *conjugate complex matrix* $\mathbf{A}^*$ has the elements

$$a^*_{\mu\nu} = \mathrm{Re}\,(a_{\mu\nu}) - \mathrm{i}\,\mathrm{Im}\,(a_{\mu\nu}). \tag{4.2c}$$

3. Transposed Matrices $\mathbf{A}^{\mathrm{T}}$

Changing the rows and columns of a matrix $\mathbf{A}$ of size (m, n) gives the *transposed matrix* $\mathbf{A}^{\mathrm{T}}$. This has the size (n, m) and

$$(a_{\nu\mu})^{\mathrm{T}} = (a_{\mu\nu}) \tag{4.3}$$

is valid.

4. Adjoint Matrices

The *adjoint matrix* $\mathbf{A}^{\mathrm{H}}$ of a complex matrix $\mathbf{A}$ is the transpose of its conjugate complex matrix $\mathbf{A}^*$ (which can not be confused with the adjoint matrix $\mathbf{A}_{adj}$, see 4.2.2, p. 278):

$$\mathbf{A}^{\mathrm{H}} = (\mathbf{A}^*)^{\mathrm{T}}. \tag{4.4}$$

5. Zero Matrix

A matrix $\mathbf{0}$ is called a *zero matrix* if it has only zero elements:

$$\mathbf{0} = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix}. \tag{4.5}$$

I.N. Bronshtein et al., *Handbook of Mathematics*,
DOI 10.1007/978-3-662-46221-8_4

4.1.2 Square Matrices

1. Definition

Square matrices have the same number of rows and columns, i.e., $m = n$:

$$\mathbf{A} = \mathbf{A}_{(n,n)} = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix}. \tag{4.6}$$

The elements $a_{\mu\nu}$ of a matrix $\mathbf{A}$ in the diagonal from the left upper corner to the right lower one are the *elements of the main diagonal.* They are denoted by $a_{11}, a_{22}, \ldots, a_{nn}$, i.e., they are all the elements $a_{\mu\nu}$ with $\mu = \nu$.

2. Diagonal Matrices

A square matrix $\mathbf{D}$ is called a *diagonal matrix* if all of the non-diagonal elements are equal to zero:

$$a_{\mu\nu} = 0 \quad \text{for} \quad \mu \neq \nu : \quad \mathbf{D} = \begin{pmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & a_{nn} \end{pmatrix} = \begin{pmatrix} a_{11} & & \mathbf{O} \\ & a_{22} & \\ & & \ddots & \\ \mathbf{O} & & & a_{nn} \end{pmatrix}. \tag{4.7}$$

3. Scalar Matrix

A diagonal matrix $\mathbf{S}$ is called a *scalar matrix* if all the diagonal elements are the same real or complex number c:

$$\mathbf{S} = \begin{pmatrix} c & 0 & \cdots & 0 \\ 0 & c & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & c \end{pmatrix}. \tag{4.8}$$

4. Trace or Spur of a Matrix

For a square matrix, the *trace* or *spur* of the matrix is defined as the sum of the main diagonal elements:

$$\operatorname{Tr}(\mathbf{A}) = a_{11} + a_{22} + \ldots + a_{nn} = \sum_{\mu=1}^{n} a_{\mu\mu}. \tag{4.9}$$

5. Symmetric Matrices

A square matrix $\mathbf{A}$ is *symmetric* if it is equal to its own transpose:

$$\mathbf{A} = \mathbf{A}^{\mathrm{T}}. \tag{4.10}$$

For the elements lying in symmetric positions with respect to the main diagonal

$$a_{\mu\nu} = a_{\nu\mu} \tag{4.11}$$

is valid.

6. Normal Matrices

satisfy the equality

$$\mathbf{A}^{\mathrm{H}}\mathbf{A} = \mathbf{A}\mathbf{A}^{\mathrm{H}}. \tag{4.12}$$

(For the product of matrices see 4.1.4, p. 272.)

7. Antisymmetric or Skew-Symmetric Matrices

are the square matrices $\mathbf{A}$ with the property:

$$\mathbf{A} = -\mathbf{A}^{\mathrm{T}}. \tag{4.13a}$$

For the elements $a_{\mu\nu}$ of an antisymmetric matrix the equalities

$$a_{\mu\nu} = -a_{\nu\mu}, \quad a_{\mu\mu} = 0 \tag{4.13b}$$

are valid, so the trace of an antisymmetric matrix vanishes:

$$\operatorname{Tr}(\mathbf{A}) = 0. \tag{4.13c}$$

The elements lying in symmetric positions with respect to the main diagonal differ from each other only in sign.
Every square matrix $\mathbf{A}$ can be decomposed into the sum of a symmetric matrix $\mathbf{A_s}$ and an antisymmetric matrix $\mathbf{A_{as}}$:

$$\mathbf{A} = \mathbf{A_s} + \mathbf{A_{as}} \qquad \text{with} \qquad \mathbf{A_s} = \frac{1}{2}(\mathbf{A} + \mathbf{A}^{\mathrm{T}}); \ \mathbf{A_{as}} = \frac{1}{2}(\mathbf{A} - \mathbf{A}^{\mathrm{T}}). \tag{4.13d}$$

8. Hermitian Matrices or Self-Adjoint Matrices

are square matrices $\mathbf{A}$ equal to their own adjoints :

$$\mathbf{A} = (\mathbf{A}^*)^{\mathrm{T}} = \mathbf{A}^{\mathrm{H}}. \tag{4.14}$$

Over the real numbers the concepts of symmetric and Hermitian matrices are the same. The determinant of a Hermitian matrix is real.

9. Anti-Hermitian or Skew-Hermitian Matrices

are the square matrices equal to their negative adjoints:

$$\mathbf{A} = -(\mathbf{A}^*)^{\mathrm{T}} = -\mathbf{A}^{\mathrm{H}}. \tag{4.15a}$$

For the elements $a_{\mu\nu}$ and the trace of an anti-Hermitian matrix the equalities

$$a_{\mu\nu} = -a^*_{\mu\nu}, \ a_{\mu\mu} = 0; \ \mathrm{Tr}\,(\mathbf{A}) = 0 \tag{4.15b}$$

are valid. Every square matrix $\mathbf{A}$ can be decomposed into a sum of a Hermitian matrix $\mathbf{A_h}$ and an anti-Hermitian matrix $\mathbf{A_{ah}}$:

$$\mathbf{A} = \mathbf{A}_{\mathrm{h}} + \mathbf{A}_{\mathrm{ah}} \quad \text{with} \quad \mathbf{A}_{\mathrm{h}} = \frac{1}{2}(\mathbf{A} + \mathbf{A}^{\mathrm{H}}), \ \mathbf{A}_{\mathrm{ah}} = \frac{1}{2}(\mathbf{A} - \mathbf{A}^{\mathrm{H}}). \tag{4.15c}$$

10. Identity Matrix I

is a diagonal matrix such that every diagonal element is equal to 1 and all of the non-diagonal elements are equal to zero:

$$\mathbf{I} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} = (\delta_{\mu\nu}) \qquad \text{with} \qquad \delta_{\mu\nu} = \begin{cases} 0 \text{ for } \mu \neq \nu, \\ 1 \text{ for } \mu = \nu. \end{cases} \tag{4.16}$$

The symbol $\delta_{\mu\nu}$ is called the *Kronecker symbol*.

11. Triangular Matrix

1. Upper Triangular Matrix, U, is a square matrix such that all the elements under the main diagonal are equal to zero:

$$\mathbf{R} = (r_{\mu\nu}) \quad \text{with} \quad r_{\mu\nu} = 0 \quad \text{for} \quad \mu > \nu. \tag{4.17}$$

2. Lower Triangular Matrix, L, is a square matrix such that all the elements above the main diagonal are equal to zero:

$$\mathbf{L} = (l_{\mu\nu}) \quad \text{with} \quad l_{\mu\nu} = 0 \quad \text{for} \quad \mu < \nu. \tag{4.18}$$

4.1.3 Vectors

Matrices of size $(n, 1)$ are one-column matrices or *column vectors* of dimension n. Matrices of size $(1, n)$ are one-row matrices or *row vectors* of dimension n:

$$\textbf{Column Vector: } \underline{\mathbf{a}} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix}, \tag{4.19a}$$

$$\textbf{Row Vector: } \underline{\mathbf{a}}^{\mathrm{T}} = (a_1, a_2, \ldots, a_n). \tag{4.19b}$$

By transposing, a column vector is changed into a row vector and conversely. A row or column vector

of dimension n can determine a point in the n dimensional Euclidean space $\mathbb{R}^n$.
The zero vector is denoted by $\underline{\mathbf{0}}$ or $\underline{\mathbf{0}}^{\mathrm{T}}$ respectively.

4.1.4 Arithmetical Operations with Matrices

1. Equality of Matrices

Two matrices $\mathbf{A} = (a_{\mu\nu})$ and $\mathbf{B} = (b_{\mu\nu})$ are equal if they have the same size and the corresponding elements are equal:

$$\mathbf{A} = \mathbf{B}, \quad \text{when} \quad a_{\mu\nu} = b_{\mu\nu} \quad \text{for} \quad \mu = 1, \ldots, m;\ \nu = 1, \ldots, n. \tag{4.20}$$

2. Addition and Subtraction

Matrices can be added or subtracted only if they have the same size. The sum/difference of two matrices is done by adding/subtracting the corresponding elements:

$$\mathbf{A} \pm \mathbf{B} = (a_{\mu\nu}) \pm (b_{\mu\nu}) = (a_{\mu\nu} \pm b_{\mu\nu}). \tag{4.21a}$$

■ $\begin{pmatrix} 1 & 3 & 7 \\ 2 & -1 & 4 \end{pmatrix} + \begin{pmatrix} 3 & -5 & 0 \\ 2 & 1 & 4 \end{pmatrix} = \begin{pmatrix} 4 & -2 & 7 \\ 4 & 0 & 8 \end{pmatrix}$.

For the addition of matrices the commutative law and the associative law are valid:

a) Commutative Law: $\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$. (4.21b)

b) Associative Law: $(\mathbf{A} + \mathbf{B}) + \mathbf{C} = \mathbf{A} + (\mathbf{B} + \mathbf{C})$. (4.21c)

3. Multiplication of a Matrix by a Number

A matrix $\mathbf{A}$ of size (m, n) is multiplied by a real or complex number α by multiplying every element of $\mathbf{A}$ by α:

$$\alpha\mathbf{A} = \alpha\,(a_{\mu\nu}) = (\alpha a_{\mu\nu}). \tag{4.22a}$$

■ $3\begin{pmatrix} 1 & 3 & 7 \\ 0 & -1 & 4 \end{pmatrix} = \begin{pmatrix} 3 & 9 & 21 \\ 0 & -3 & 12 \end{pmatrix}$.

From (4.22a) it is obvious that one can factor out a constant multiplier contained by every element of a matrix. For the multiplication of a matrix by a scalar the *commutative, associative and distributive laws for multiplication* are valid:

a) Commutative Law: $\alpha\mathbf{A} = \mathbf{A}\alpha$; (4.22b)

b) Associative Law: $\alpha(\beta\mathbf{A}) = (\alpha\beta)\mathbf{A}$; (4.22c)

c) Distributive Law: $(\alpha \pm \beta)\mathbf{A} = \alpha\mathbf{A} \pm \beta\mathbf{A}$; $\quad \alpha(\mathbf{A} \pm \mathbf{B}) = \alpha\mathbf{A} \pm \alpha\mathbf{B}$. (4.22d)

4. Division of a Matrix by a Number

The *division of a matrix by a scalar* $\gamma \neq 0$ is the same as multiplication by $\alpha = 1/\gamma$.

5. Multiplication of Two Matrices

1. The Product A B of two matrices $\mathbf{A}$ and $\mathbf{B}$ can be calculated only if the number of columns of the factor $\mathbf{A}$ on the left-hand side is equal to the number of rows of the factor $\mathbf{B}$ on the right-hand side. If $\mathbf{A}$ is a matrix of size (m, n), then the matrix $\mathbf{B}$ must have size (n, p), and the product $\mathbf{A}\,\mathbf{B}$ is a matrix $\mathbf{C} = (c_{\mu\lambda})$ of size (m, p). The element $c_{\mu\lambda}$ is equal to the scalar product of the μ-th row of the factor $\mathbf{A}$ on the left with the λ-th column of the factor $\mathbf{B}$ on the right:

$$\mathbf{A}\,\mathbf{B} = \Big(\sum_{\nu=1}^{n} a_{\mu\nu} b_{\nu\lambda}\Big) = (c_{\mu\lambda}) = \mathbf{C} \qquad (\mu = 1, 2, \ldots, m;\ \lambda = 1, 2, \ldots, p). \tag{4.23}$$

■ $\mathbf{A} = \begin{pmatrix} 1 & 3 & 7 \\ \boxed{2 \ \ -1 \ \ 4} \\ -1 & 0 & 1 \end{pmatrix}$, $\mathbf{B} = \begin{pmatrix} 3 & \boxed{2} \\ -5 & \boxed{1} \\ 0 & \boxed{3} \end{pmatrix}$. The element c_{22} of the product matrix C in accordance with (4.23) is $c_{22} = 2 \cdot 2 - 1 \cdot 1 + 4 \cdot 3 = 15$.

2. Inequality of Matrix Products Even if both products $\mathbf{A}\,\mathbf{B}$ and $\mathbf{B}\,\mathbf{A}$ exist, usually $\mathbf{A}\,\mathbf{B} \neq \mathbf{B}\,\mathbf{A}$, i.e., in general the commutative law for multiplication is not valid. If the equality $\mathbf{A}\,\mathbf{B} = \mathbf{B}\,\mathbf{A}$ holds, then one says that the matrices $\mathbf{A}$ and $\mathbf{B}$ are commutable or commute with each other.

3. Falk Scheme Multiplication of matrices $\mathbf{A}\,\mathbf{B} = \mathbf{C}$ can be performed using the *Falk scheme* **(Fig. 4.1)**. The element $c_{\mu\lambda}$ of the product matrix $\mathbf{C}$ appears exactly at the intersection point of the μ-th row of $\mathbf{A}$ with the λ-th column of $\mathbf{B}$.

■ Multiplication of the matrices $\mathbf{A}_{(3,3)}$ and $\mathbf{B}_{(3,2)}$ is shown in **Fig. 4.2** using the Falk scheme.

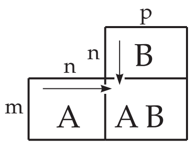


Figure 4.1

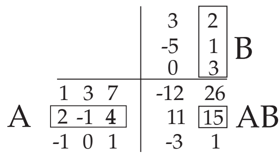


Figure 4.2

4. Multiplication of the Matrices $\mathbf{K_1}$ and $\mathbf{K_2}$ with Complex Elements For multiplication of two matrices with complex elements can be used their decompositions into real and imaginary parts according to (4.2b): $\mathbf{K}_1 = \mathbf{A}_1 + \mathrm{i}\mathbf{B}_1$, $\mathbf{K}_2 = \mathbf{A}_2 + \mathrm{i}\mathbf{B}_2$. Here $\mathbf{A_1}, \mathbf{A_2}, \mathbf{B_1}, \mathbf{B_2}$ are real matrices. After this decomposition, the multiplication results in a sum of matrices whose terms are products of real matrices.

■ $(\mathbf{A} + \mathrm{i}\,\mathbf{B})(\mathbf{A} - \mathrm{i}\,\mathbf{B}) = \mathbf{A}^2 + \mathbf{B}^2 + \mathrm{i}\,(\mathbf{BA} - \mathbf{AB})$ (Powers of Matrices see 4.1.5,**8.**, p. 276). Of course when multiplying these matrices it must be considered that the commutative law for multiplication is not valid in general, i.e., the matrices $\mathbf{A}$ and $\mathbf{B}$ do not usually commute with each other.

6. Scalar and Dyadic Product of Two Vectors

If the vectors $\underline{\mathbf{a}}$ and $\underline{\mathbf{b}}$ are considered as one-row and one-column matrices, respectively, then there are two possibilities to multiply them according to the rules of matrix multiplication:
If $\underline{\mathbf{a}}$ has size $(1, n)$ and $\underline{\mathbf{b}}$ has size $(n, 1)$ then their product has size $(1, 1)$, i.e. it is a number. It is called the *scalar product* of two vectors. If conversely, $\underline{\mathbf{a}}$ has size $(n, 1)$ and $\underline{\mathbf{b}}$ has size $(1, m)$, then the product has size (n, m), i.e., it is a matrix. This matrix is called the *dyadic product* of the two vectors.

1. Scalar Product of Two Vectors The scalar product of a row vector $\underline{\mathbf{a}}^{\mathrm{T}} = (a_1, a_2, \ldots, a_n)$ with a column vector $\underline{\mathbf{b}} = (b_1, b_2, \ldots, b_n)^{\mathrm{T}}$ – both having n elements – is defined as the number

$$\underline{\mathbf{a}}^{\mathrm{T}}\underline{\mathbf{b}} = \underline{\mathbf{b}}^{\mathrm{T}}\underline{\mathbf{a}} = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n = \sum_{\mu=1}^{n} a_\mu b_\mu. \tag{4.24}$$

The commutative law for multiplication is not valid for a product of vectors in general, so one must keep the exact order of $\underline{\mathbf{a}}^{\mathrm{T}}$ and $\underline{\mathbf{b}}$. If the order of multiplication is reversed, then the product $\underline{\mathbf{b}}\underline{\mathbf{a}}^{\mathrm{T}}$ is a dyadic product.

2. Dyadic Product or Tensor Product of Two Vectors The *dyadic product* of a column vector $\underline{\mathbf{a}} = (a_1, a_2, \ldots, a_n)^{\mathrm{T}}$ of dimension n with a row vector $\underline{\mathbf{b}}^{\mathrm{T}} = (b_1, b_2, \ldots, b_m)$ of dimension m is defined as the following matrix:

$$\underline{\mathbf{a}}\underline{\mathbf{b}}^{\mathrm{T}} = \begin{pmatrix} a_1 b_1 & a_1 b_2 & \cdots & a_1 b_m \\ a_2 b_1 & a_2 b_2 & \cdots & a_2 b_m \\ \vdots & \vdots & \vdots & \vdots \\ a_n b_1 & a_n b_2 & \cdots & a_n b_m \end{pmatrix} \tag{4.25}$$

of size (n, m). Also here the commutative law for multiplication is not valid in general.

3. Hints on the Notion of Vector Products of Two Vectors In the domain of multi-vectors or alternating tensors there is a so-called outer product whose three-dimensional version is the well-

known *vector product* or *cross product* (see 3.5.1.5, **2.**, p. 184 ff). In this book the outer product of multi-vectors of higher rank is not discussed.

7. Rank of a Matrix

1. Definition In a matrix $\mathbf{A}$ the maximum number r of linearly independent column vectors is equal to the maximum number of linearly independent row vectors. This number r is called the *rank of the matrix* and it is denoted by $\operatorname{rank}(\mathbf{A}) = r$.

2. Statements about the Rank of a Matrix

a) Because in a vector space of dimension m there exist no more than m linearly independent m-dimensional row or column vectors (see 5.3.8.2, p. 366), the rank r of a matrix $\mathbf{A}$ of size (m, n) cannot be greater, than the smaller of m and n:

$$\operatorname{rank}(\mathbf{A}_{(m,n)}) = r \le \min(m, n). \tag{4.26a}$$

b) A square matrix $\mathbf{A}_{(n,n)}$ is called a *regular matrix* if

$$\operatorname{rank}(\mathbf{A}_{(n,n)}) = r = n. \tag{4.26b}$$

A square matrix of size (n, n) is regular if and only if its determinant differs from zero, i.e., $\det \mathbf{A} \neq 0$ (see 4.2.2, **3.**, p. 279). Otherwise it is a *singular matrix*.

c) Consequently for the rank of a singular square matrix $\mathbf{A}_{(n,n)}$, i.e., $\det \mathbf{A} = 0$

$$\operatorname{rank}(\mathbf{A}_{(n,n)}) = r < n \tag{4.26c}$$

is valid.

d) The rank of the zero matrix $\mathbf{0}$ is equal to zero:

$$\operatorname{rank}(\mathbf{0}) = r = 0. \tag{4.26d}$$

e) The rank of the sum and product of matrices satisfies the relations

$$|\operatorname{rank}(\mathbf{A}) - \operatorname{rank}(\mathbf{B})| \le \operatorname{rank}(\mathbf{A} + \mathbf{B}) \le \operatorname{rank}(\mathbf{A}) + \operatorname{rank}(\mathbf{B}), \tag{4.26e}$$

$$\operatorname{rank}(\mathbf{AB}) \le \min(\operatorname{rank}(\mathbf{A}), \operatorname{rank}(\mathbf{B})). \tag{4.26f}$$

3. Rules to Determine the Rank Elementary transformations do not change the rank of matrices. *Elementary transformations* in this relation are:

a) Interchanging two columns or two rows.

b) Multiplication of a row or column by a number.

c) Addition of a row to another row or a column to an other column.

In order to determine their ranks every matrix can be transformed by appropriate linear combinations of rows into a form such that in the μ-th row ($\mu = 2, 3, \ldots, m$), at least the first $\mu - 1$ elements are equal to zero (the principle of Gauss algorithm, see 4.5.2.4, p. 312). The number of row vectors different from the zero vector in the transformed matrix is equal to the rank r of the matrix.

8. Inverse Matrix

For a regular matrix $\mathbf{A} = (a_{\mu\nu})$ there is always an *inverse matrix* $\mathbf{A}^{-1}$ (with respect to multiplication), i.e., the multiplication of a matrix by its inverse yields the identity matrix:

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}. \tag{4.27a}$$

The elements of $\mathbf{A}^{-1} = (\beta_{\mu\nu})$ are

$$\beta_{\mu\nu} = \frac{\mathbf{A}_{\nu\mu}}{\det \mathbf{A}}, \tag{4.27b}$$

where $\mathbf{A}_{\nu\mu}$ is the cofactor belonging to the $a_{\nu\mu}$ element of the matrix $\mathbf{A}$ (see 4.2.1, **1.**, p. 278). For a practical calculation of $\mathbf{A}^{-1}$ the method given in 4.2.2, **2.**, p. 278 should be used. In the case of a matrix of size $(2, 2)$ holds:

$$\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \qquad \mathbf{A}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}. \tag{4.28}$$

Remark: Why not define division among matrices but instead use the inverse for calculations? This is connected to the fact that division cannot be defined uniquely. The solutions of the equations

$$\begin{matrix} \mathbf{B}\mathbf{X_1} = \mathbf{A} \\ \mathbf{X_2}\mathbf{B} = \mathbf{A} \end{matrix} \quad (\mathbf{B} \text{ regular}), \qquad \begin{matrix} \mathbf{X_1} = \mathbf{B^{-1}}\mathbf{A} \\ \mathbf{X_2} = \mathbf{A}\mathbf{B^{-1}} \end{matrix} \tag{4.29}$$

are in general different.

9. Orthogonal Matrices

If the relation

$$\mathbf{A}^{\mathrm{T}} = \mathbf{A^{-1}} \qquad \text{or} \qquad \mathbf{A}\mathbf{A}^{\mathrm{T}} = \mathbf{A}^{\mathrm{T}}\mathbf{A} = \mathbf{I} \tag{4.30}$$

holds for a square matrix $\mathbf{A}$, then it is called an *orthogonal matrix*, i.e., the scalar product of a row and the transpose of another one, or the scalar product of the transpose of a column and another one are zero, while the scalar product of a row with its own transpose or of the transpose of a column with itself are equal to one.

Orthogonal matrices have the following properties:

a) The transpose and the inverse of an orthogonal matrix $\mathbf{A}$ are also orthogonal; furthermore, the determinant is

$$\det \mathbf{A} = \pm 1. \tag{4.31}$$

b) Products of orthogonal matrices are also orthogonal.

■ The *rotation matrix* $\mathbf{D}$, which is used to describe the rotation of a coordinate system, and whose elements are the direction cosines of the new direction of axes (see 3.5.3.3,**2.** p. 212), is also an orthogonal matrix.

10. Unitary Matrix

If for a matrix $\mathbf{A}$ with complex elements

$$(\mathbf{A}^*)^{\mathrm{T}} = \mathbf{A^{-1}} \qquad \text{or} \qquad \mathbf{A}(\mathbf{A}^*)^{\mathrm{T}} = (\mathbf{A}^*)^{\mathrm{T}}\mathbf{A} = \mathbf{I} \tag{4.32}$$

holds it is called a *unitary matrix*. In the real case unitary and orthogonal matrices are the same.

4.1.5 Rules of Calculation for Matrices

The following rules are valid of course only in the case when the operations can be performed, for instance the identity matrix $\mathbf{I}$ always has a size corresponding to the requirements of the given operation.

1. Multiplication of a Matrix by the Identity Matrix

is also called the *identical transformation*:

$$\mathbf{A}\mathbf{I} = \mathbf{I}\mathbf{A} = \mathbf{A}. \tag{4.33}$$

(This does not mean that the commutative law is valid in general, because the sizes of the matrix $\mathbf{I}$ on the left- and on the right-hand side may be different.)

2. Multiplication of a Square Matrix A by a Scalar Matrix S

or by the identity matrix $\mathbf{I}$ is commutative

$$\mathbf{A}\mathbf{S} = \mathbf{S}\mathbf{A} = c\mathbf{A} \quad \text{with } \mathbf{S} \text{ given in (4.8)}, \quad (4.34\text{a}) \qquad \mathbf{A}\mathbf{I} = \mathbf{I}\mathbf{A} = \mathbf{A}. \quad (4.34\text{b})$$

3. Multiplication of a Matrix A by the Zero Matrix 0

results in the zero matrix:

$$\mathbf{A}\mathbf{0} = \mathbf{0}, \qquad \mathbf{0}\mathbf{A} = \mathbf{0}. \tag{4.35}$$

(The zero matrices above may have different sizes.) The converse statement is not true in general, i.e., from $\mathbf{A}\mathbf{B} = \mathbf{0}$ it does not follow that $\mathbf{A} = \mathbf{0}$ or $\mathbf{B} = \mathbf{0}$.

4. Vanishing Product of Two Matrices

The product of two matrices $\mathbf{A}$ and $\mathbf{B}$ can be the zero matrix even if neither of them is a zero matrix:

$$\mathbf{A}\,\mathbf{B} = \mathbf{0} \text{ or } \mathbf{B}\,\mathbf{A} = \mathbf{0} \text{ or both, although } \mathbf{A} \neq \mathbf{0},\ \mathbf{B} \neq \mathbf{0}. \tag{4.36}$$

■ $$\begin{array}{cc|cc} & & 1 & 1 \\ & & 0 & 0 \\ \hline 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{array}.$$

5. Multiplication of Three Matrices

$$(\mathbf{A}\,\mathbf{B})\mathbf{C} = \mathbf{A}\,(\mathbf{B}\,\mathbf{C}) \tag{4.37}$$

i.e., the associative law of multiplication is valid.

6. Transposing of a Sum or a Product of Two Matrices

$$(\mathbf{A}+\mathbf{B})^{\mathrm{T}} = \mathbf{A}^{\mathrm{T}} + \mathbf{B}^{\mathrm{T}}, \qquad (\mathbf{AB})^{\mathrm{T}} = \mathbf{B}^{\mathrm{T}}\,\mathbf{A}^{\mathrm{T}}, \qquad (\mathbf{A}^{\mathrm{T}})^{\mathrm{T}} = \mathbf{A}. \tag{4.38a}$$

For square invertible matrices $\mathbf{A}_{(n,n)}$:

$$(\mathbf{A}^{\mathrm{T}})^{-\mathbf{1}} = (\mathbf{A}^{-\mathbf{1}})^{\mathrm{T}} \tag{4.38b}$$

holds.

7. Inverse of a Product of Two Matrices

$$(\mathbf{A}\,\mathbf{B})^{-\mathbf{1}} = \mathbf{B}^{-\mathbf{1}}\,\mathbf{A}^{-\mathbf{1}}. \tag{4.39}$$

8. Powers of Matrices

$$\mathbf{A}^p = \underbrace{\mathbf{A}\,\mathbf{A}\ldots\mathbf{A}}_{p \text{ factors}} \quad \text{with } p>0, \text{ integer}, \tag{4.40a}$$

$$\mathbf{A}^0 = \mathbf{I} \qquad (\det\mathbf{A} \neq 0)\,, \tag{4.40b}$$

$$\mathbf{A}^{-p} = (\mathbf{A}^{-\mathbf{1}})^p \qquad (p>0, \text{ integer}; \det\mathbf{A} \neq 0), \tag{4.40c}$$

$$\mathbf{A}^{p+q} = \mathbf{A}^p\mathbf{A}^q \qquad (p, q \text{ integer}). \tag{4.40d}$$

9. Kronecker Product

The *Kronecker product of two matrices* $\mathbf{A} = (a_{\mu\nu})$ of the type (m,n) and $\mathbf{B} = (b_{\mu\nu})$ of the type (p,r) is defined as the rule

$$\mathbf{A} \otimes \mathbf{B} = (a_{\mu\nu}\,\mathbf{B})\,. \tag{4.41}$$

The result is a new matrix of type $(m\cdot p\,,\ n\cdot r)$, arising from the multiplication of every element of $\mathbf{A}$ by the matrix $\mathbf{B}$.

■ $\mathbf{A} = \begin{pmatrix} 3 & -5 & 0 \\ 2 & 1 & 3 \end{pmatrix}$ of type $(2,3)$, $\mathbf{B} = \begin{pmatrix} 1 & 3 \\ 2 & -1 \end{pmatrix}$ of type $(2,2)$.

$$\mathbf{A} \bigotimes \mathbf{B} = \begin{pmatrix} 3\cdot\begin{pmatrix} 1 & 3 \\ 2 & -1 \end{pmatrix} & -5\cdot\begin{pmatrix} 1 & 3 \\ 2 & -1 \end{pmatrix} & 0\cdot\begin{pmatrix} 1 & 3 \\ 2 & -1 \end{pmatrix} \\ 2\cdot\begin{pmatrix} 1 & 3 \\ 2 & -1 \end{pmatrix} & 1\cdot\begin{pmatrix} 1 & 3 \\ 2 & -1 \end{pmatrix} & 3\cdot\begin{pmatrix} 1 & 3 \\ 2 & -1 \end{pmatrix} \end{pmatrix} = \begin{pmatrix} 3 & 9 & -5 & -15 & 0 & 0 \\ 6 & -3 & -10 & 5 & 0 & 0 \\ 2 & 6 & 1 & 3 & 3 & 9 \\ 4 & -2 & 2 & -1 & 6 & -3 \end{pmatrix},$$

gives a matrix of type $(4,6)$.

For the transpose and the trace are valid the equalities:

$$(\mathbf{A} \otimes \mathbf{B})^{\mathrm{T}} = \mathbf{A}^{\mathrm{T}} \otimes \mathbf{B}^{\mathrm{T}}\,, \tag{4.42}$$

$$\mathrm{Tr}\,(\mathbf{A} \otimes \mathbf{B}) = \mathrm{Tr}\,(\mathbf{A}) \cdot \mathrm{Tr}(\mathbf{B})\,. \tag{4.43}$$

10. Differentiation of a Matrix

If a matrix $\mathbf{A} = \mathbf{A}(t) = (a_{\mu\nu}(t))$ has differentiable elements $a_{\mu\nu}(t)$ of a parameter t then its derivative with respect to t is given as

$$\frac{d\mathbf{A}}{dt} = \left(\frac{da_{\mu\nu}(t)}{dt}\right) = (a'_{\mu\nu}(t))\,. \tag{4.44}$$

4.1.6 Vector and Matrix Norms

The norm of a vector or of a matrix can be considered as a generalization of the absolute value (magnitude) of numbers. Therefore a real number is assigned as $||x||$ (*Norm* $\underline{\mathbf{x}}$) to the vector $\underline{\mathbf{x}}$ or as $||\mathbf{A}||$

(*Norm* $\mathbf{A}$) to a matrix $\mathbf{A}$. These numbers must satisfy the norm axioms (see 12.3.1.1, p. 669). For vectors $\underline{\mathbf{x}} \in \mathbb{R}^n$ they are:

1. $||\underline{\mathbf{x}}|| \geq 0$ for every $\underline{\mathbf{x}}$; $||\underline{\mathbf{x}}|| = 0$ if and only if $\underline{\mathbf{x}} = 0$. (4.45)

2. $||\lambda\underline{\mathbf{x}}|| = |\lambda|\,||\underline{\mathbf{x}}||$ for every $\underline{\mathbf{x}}$ and every real number λ. (4.46)

3. $||\underline{\mathbf{x}} + \underline{\mathbf{y}}|| \leq ||\underline{\mathbf{x}}|| + ||\underline{\mathbf{y}}||$ for every $\underline{\mathbf{x}}$ and $\underline{\mathbf{y}}$ (triangle inequality) (see also 3.5.1.1, **1.**, p. 182). (4.47)

There are many different ways to define norms for vectors and matrices. But for practical reasons it is better to define a matrix norm $||\mathbf{A}||$ and a vector norm $||\underline{\mathbf{x}}||$ so that they might satisfy the inequality

$$||\mathbf{A}\underline{\mathbf{x}}|| \leq ||\mathbf{A}||\,||\underline{\mathbf{x}}||. \tag{4.48}$$

This inequality is very useful for error estimations. If the matrix and vector norms satisfy this inequality, then one says that they are *consistent with each other*. If there is a non-zero vector $\underline{\mathbf{x}}$ for every $\mathbf{A}$ such that the equality holds in (4.48), then one says *the matrix norm* $||\mathbf{A}||$ is the *subordinate to the vector norm* $||\underline{\mathbf{x}}||$.

4.1.6.1 Vector Norms

If $\underline{\mathbf{x}} = (x_1, x_2, \ldots, x_n)^{\mathrm{T}}$ is a real vector of n dimensions, i.e., $\underline{\mathbf{x}} \in \mathbb{R}^n$, then the most often used vector norms are:

1. Euclidean Norm

$$||\underline{\mathbf{x}}|| = ||\underline{\mathbf{x}}||_2 := \sqrt{\sum_{i=1}^{n} x_i^2}\,. \tag{4.49}$$

2. Supremum or Uniform Norm

$$||\underline{\mathbf{x}}|| = ||\underline{\mathbf{x}}||_\infty := \max_{1\leq i\leq n}|x_i|. \tag{4.50}$$

3. Sum Norm

$$||\underline{\mathbf{x}}|| = ||\underline{\mathbf{x}}||_1 := \sum_{i=1}^{n} |\mathbf{x}_i|. \tag{4.51}$$

■ In $\mathbb{R}^3$, in elementary vector calculus $||\underline{\mathbf{x}}||_2$ is considered as the magnitude of the vector $\underline{\mathbf{x}}$. The magnitude $|x| = ||\underline{\mathbf{x}}||_2$ gives the length of the vector $\underline{\mathbf{x}}$.

4.1.6.2 Matrix Norms

1. Spectral Norm for Real Matrices

$$||\mathbf{A}|| = ||\mathbf{A}||_2 := \sqrt{\lambda_{\max}(\mathbf{A}^{\mathrm{T}}\mathbf{A})}. \tag{4.52}$$

Here $\lambda_{\max}(\mathbf{A}^{\mathrm{T}}\mathbf{A})$ denotes the greatest eigenvalue (see 4.6.1, p. 314) of the matrix $\mathbf{A}^{\mathrm{T}}\mathbf{A}$.

2. Row-Sum Norm

$$||\mathbf{A}|| = ||\mathbf{A}||_\infty := \max_{1\leq i\leq n} \sum_{j=1}^{n} |a_{ij}|. \tag{4.53}$$

3. Column-Sum Norm

$$||\mathbf{A}|| = ||\mathbf{A}||_1 := \max_{1\leq j\leq n} \sum_{i=1}^{n} |a_{ij}|. \tag{4.54}$$

It can be proved that the matrix norm (4.52) is the subordinate norm to the vector norm (4.49). The same is true for (4.53) and (4.50), and for (4.54) and (4.51).

4.2 Determinants

4.2.1 Definitions

1. Determinants

Determinants are real or complex numbers uniquely associated with square matrices. The *determinant* of order n associated with the (n, n) matrix $\mathbf{A} = (a_{\mu\nu})$,

$$\mathrm{D} = \det \mathbf{A} = \det (a_{\mu\nu}) = \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{vmatrix}, \tag{4.55}$$

is calculated in a recursive way using the *Laplace expansion rule*:

$$\det \mathbf{A} = \sum_{\nu=1}^{n} a_{\mu\nu} A_{\mu\nu} \qquad (\mu \text{ fixed, expansion along the } \mu\text{-th row}), \tag{4.56a}$$

$$\det \mathbf{A} = \sum_{\mu=1}^{n} a_{\mu\nu} A_{\mu\nu} \qquad (\nu \text{ fixed, expansion along the } \nu\text{-th column}). \tag{4.56b}$$

Here $A_{\mu\nu}$ is the subdeterminant belonging to the element $a_{\mu\nu}$ multiplied by the sign factor $(-1)^{\mu+\nu}$. $A_{\mu\nu}$ is called the *cofactor* or *algebraic complement*.

2. Subdeterminants

The *subdeterminant* of order $(n-1)$ belonging to the element $a_{\mu\nu}$ of a determinant of order n is the determinant obtained by deleting the μ-th row and the ν-th column.

■ Expansion of a determinant of order four along the third row:

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{vmatrix} = a_{31} \begin{vmatrix} a_{12} & a_{13} & a_{14} \\ a_{22} & a_{23} & a_{24} \\ a_{42} & a_{43} & a_{44} \end{vmatrix} - a_{32} \begin{vmatrix} a_{11} & a_{13} & a_{14} \\ a_{21} & a_{23} & a_{24} \\ a_{41} & a_{43} & a_{44} \end{vmatrix} + a_{33} \begin{vmatrix} a_{11} & a_{12} & a_{14} \\ a_{21} & a_{22} & a_{24} \\ a_{41} & a_{42} & a_{44} \end{vmatrix} - a_{34} \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{41} & a_{42} & a_{43} \end{vmatrix}.$$

4.2.2 Rules of Calculation for Determinants

Because of the Laplace expansion the following statements about rows are valid also for columns.

1. Independence of the Value of a Determinant

The value of a determinant does not depend on which row was chosen.

2. Substitution of Cofactors

If during the expansion of a determinant the cofactors of a row are replaced by the cofactors of another one, then one gets zero:

$$\sum_{\nu=1}^{n} a_{\mu\nu} A_{\lambda\nu} = 0 \qquad (\mu, \lambda \text{ fixed; } \lambda \neq \mu). \tag{4.57}$$

This relation and the Laplace expansion result in

$$\mathbf{A}_{\mathrm{adj}}\, \mathbf{A} = \mathbf{A}\, \mathbf{A}_{\mathrm{adj}} = (\det \mathbf{A})\, \mathbf{I}. \tag{4.58}$$

The *adjoint matrix* of $\mathbf{A}$, which is the transpose of the matrix made from the cofactors of $\mathbf{A}$, is denoted by $\mathbf{A}_{\mathrm{adj}}$. There must not be a confusion of this adjoint matrix with the transposed conjugate of a complex matrix $\mathbf{A}^{\mathrm{H}}$ (see (4.4), p. 269). From the previous equality one gets the *inverse matrix*

$$\mathbf{A}^{-1} = \frac{1}{\det \mathbf{A}} \mathbf{A}_{\mathrm{adj}}, \tag{4.59}$$

3. Zero Value of a Determinant

A determinant is equal to zero if

a) a row contains zero elements only, or

b) two rows are equal to each other, or

c) a row is a linear combination of the others.

4. Changes and Additions

The value of the determinant does not change if

a) its rows are exchanged for its columns, i.e., *reflection* in the main diagonal does not affect the value of it:

$$\det \mathbf{A} = \det \mathbf{A}^{\mathrm{T}}, \tag{4.60}$$

b) any row is added to or subtracted from another one, or

c) a multiple of any row is added to or subtracted from another one, or

d) a linear combination of other rows is added to any row.

5. Sign on Changing Rows

If two rows are interchanged in a determinant, then the sign of the determinant changes.

6. Multiplication of a Determinant by a Number

The value of a determinant will be multiplied by α if the elements of a row are multiplied by this number. The next formula shows the difference between this and the multiplication of a matrix $\mathbf{A}$ of size (n, n) by a number α

$$\det(\alpha\mathbf{A}) = \alpha^n \det \mathbf{A}\,. \tag{4.61}$$

7. Multiplication of Two Determinants

The multiplication of two determinants can be reduced to the multiplication of their matrices:

$$(\det \mathbf{A})(\det \mathbf{B}) = \det(\mathbf{AB}). \tag{4.62}$$

Since $\det \mathbf{A} = \det \mathbf{A}^{\mathrm{T}}$ (see (4.60)), we have the equalities

$$(\det \mathbf{A})(\det \mathbf{B}) = \det(\mathbf{AB}) = \det(\mathbf{AB}^{\mathrm{T}}) = \det(\mathbf{A}^{\mathrm{T}}\mathbf{B}) = \det(\mathbf{A}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}), \tag{4.63}$$

i.e., it is permissible to take the scalar product of rows with columns, rows with rows, columns with rows or columns with columns.

8. Differentiation of a Determinant

Suppose the elements of a determinant of order n are differentiable functions of a parameter t, i.e., $a_{\mu\nu} = a_{\mu\nu}(t)$. In order to differentiate the determinant with respect to t, one differentiates one row at one time and finally one adds the n determinants.

■ For a determinant of size $(3, 3)$ follows:

$$\frac{d}{dt}\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = \begin{vmatrix} a'_{11} & a'_{12} & a'_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} + \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a'_{21} & a'_{22} & a'_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} + \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a'_{31} & a'_{32} & a'_{33} \end{vmatrix}.$$

4.2.3 Evaluation of Determinants

1. Value of a Determinant of Second Order

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{21}a_{12}. \tag{4.64}$$

2. Value of a Determinant of Third Order

The *Sarrus rule* gives a convenient scheme for the calculations, but it is valid only for determinants of order three. It is the following:

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} \begin{matrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{matrix} = a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32}$$

$$-(a_{31}a_{22}a_{13} + a_{32}a_{23}a_{11} + a_{33}a_{21}a_{12}). \qquad (4.65)$$

The first two columns are copied after the determinant, then the sum of the products of the elements along the redrawn declining segments is calculated, then the sum of the products of the elements along the dotted inclining segments is subtracted.

3. Value of a Determinant of n-th Order

By the expansion rule the calculation of the value of a determinant of order n is reduced to the evaluation of n determinants of order $(n-1)$. But for practical reasons (to reduce the number of required operations), first one transforms the determinant with the help of the rules discussed above into a form such that it contains as many zeros as possible.

■ $$\begin{vmatrix} 2 & 9 & 9 & 4 \\ 2 & -3 & 12 & 8 \\ 4 & 8 & 3 & -5 \\ 1 & 2 & 6 & 4 \end{vmatrix} \underset{\text{(rule 4)}}{=} \begin{vmatrix} 2 & 5 & 9 & 4 \\ 2 & -7 & 12 & 8 \\ 4 & 0 & 3 & -5 \\ 1 & 0 & 6 & 4 \end{vmatrix} \underset{\text{(rule 6)}}{=} 3\begin{vmatrix} 2 & 5 & 3 & 4 \\ 2 & -7 & 4 & 8 \\ 4 & 0 & 1 & -5 \\ 1 & 0 & 2 & 4 \end{vmatrix} = 3\left(-5\underbrace{\begin{vmatrix} 2 & 4 & 8 \\ 4 & 1 & -5 \\ 1 & 2 & 4 \end{vmatrix}}_{=0 \text{ (rule 3)}} -7\begin{vmatrix} 2 & 3 & 4 \\ 4 & 1 & -5 \\ 1 & 2 & 4 \end{vmatrix}\right)$$

$$= -21\underset{\text{(rule 4)}}{\begin{vmatrix} 1 & 1 & 0 \\ 4 & 1 & -5 \\ 1 & 2 & 4 \end{vmatrix}} = -21\left(\begin{vmatrix} 1 & -5 \\ 2 & 4 \end{vmatrix} - \begin{vmatrix} 4 & -5 \\ 1 & 4 \end{vmatrix}\right) = 147.$$

Remark: An especially efficient method to determine the value of a determinant of order n can be obtained by transforming it in the same way as it is done in order to determine the rank of a matrix (see 4.1.4, **7.**, p. 274), i.e., all the elements under the diagonal $a_{11}, a_{22}, \ldots, a_{nn}$ are equal to zero. Then the value of the determinant is the product of the diagonal elements of the transformed determinant.

4.3 Tensors

4.3.1 Transformation of Coordinate Systems

1. Linear Transformation

By the linear transformation

$$\underline{\tilde{\mathbf{x}}} = \mathbf{A}\underline{\mathbf{x}} \qquad \text{or} \qquad \begin{aligned} \tilde{x}_1 &= a_{11}x_1 + a_{12}x_2 + a_{13}x_3 \\ \tilde{x}_2 &= a_{21}x_1 + a_{22}x_2 + a_{23}x_3 \\ \tilde{x}_3 &= a_{31}x_1 + a_{32}x_2 + a_{33}x_3 \end{aligned} \qquad (4.66)$$

a coordinate transformation is defined in the three-dimensional space. Here x_μ and $\tilde{x}_\mu$ $(\mu = 1, 2, 3)$ are the coordinates of the same point but in different coordinate systems K and $\tilde{K}$.

2. Einstein's Summation Convention

Instead of (4.66) one can write

$$\tilde{x}_\mu = \sum_{\nu=1}^{3} a_{\mu\nu}x_\nu \qquad (\mu = 1, 2, 3) \qquad (4.67a)$$

or briefly by Einstein

$$x_\mu = a_{\mu\nu}x_\nu, \qquad (4.67b)$$

i.e., it is to calculate the sum with respect to the repeated index ν and put down the result for $\mu = 1, 2, 3$. In general, the *summation convention* means that if an index appears twice in an expression, then the expression is added for all values of this index. If an index appears only once in the expressions of an

equation, for instance μ in (4.67b), then it means that the equality is valid for all possible values of this index.

3. Rotation of a Coordinate System

If the Cartesian coordinate system $\tilde{K}$ is given by rotation of the system K, then for the transformation matrix in (4.66) $\mathbf{A} = \mathbf{D}$ is valid. Here $\mathbf{D} = (d_{\mu\nu})$ is the orthogonal *rotation matrix*. The orthogonal rotation matrix $\mathbf{D}$ has the property

$$\mathbf{D}^{-1} = \mathbf{D}^{\mathrm{T}}. \tag{4.68a}$$

The elements $d_{\mu\nu}$ of $\mathbf{D}$ are the direction cosines of the angles between the old and new coordinate axes. From the orthogonality of $\mathbf{D}$, i.e., from

$$\mathbf{D}\mathbf{D}^{\mathrm{T}} = \mathbf{I} \quad \text{and} \quad \mathbf{D}^{\mathrm{T}}\mathbf{D} = \mathbf{I}, \tag{4.68b}$$

it follows that

$$\sum_{i=1}^{3} d_{\mu i} d_{\nu i} = \delta_{\mu\nu}, \qquad \sum_{k=1}^{3} d_{k\mu} d_{k\nu} = \delta_{\mu\nu} \qquad (\mu, \nu = 1, 2, 3). \tag{4.68c}$$

The equalities in (4.68c) show that the row and column vectors of the matrix $\mathbf{D}$ are *orthonormalized*, because $\delta_{\mu\nu}$ is the Kronecker symbol (see 4.1.2, **10.**, p. 271).
The elements $d_{\mu\nu}$ of the rotation matrix can be determined by the Cardan angles (see 3.5.3.5, p. 214) or Euler angles (see 3.5.3.6, p. 215). For rotation in the plane see 3.5.2.2,**2.**, p. 191; in space see 3.5.3.3, p. 213.

4.3.2 Tensors in Cartesian Coordinates

1. Definition

A mathematical or a physical quantity $\boldsymbol{T}$ can be described in a Cartesian coordinate system K by 3^n elements $t_{ij\cdots m}$, the so-called translation invariants. Here the number of indices $i, j, \ldots, m$ is exactly equal to n $(n \geq 0)$. The indices are ordered, and every of them takes the values 1, 2 and 3.

If under a coordinate transformation from K to $\tilde{K}$ for the elements $t_{ij\cdots m}$ according to (4.66)

$$\tilde{t}_{\mu\nu\cdots\rho} = \sum_{i=1}^{3}\sum_{j=1}^{3}\cdots\sum_{m=1}^{3} a_{\mu i} a_{\nu j} \cdots a_{\rho m} t_{ij\cdots m}\,, \tag{4.69}$$

is valid, then $\boldsymbol{T}$ is called a *tensor of rank* n, and the elements $t_{ij\cdots m}$ (mostly numbers) with ordered indices are the *components of the tensor* $\boldsymbol{T}$.

2. Tensor of Rank 0

A *tensor of rank zero* has only one component, i.e., it is a scalar. Because its value is the same in every coordinate system, one talks about the *invariance of scalars* or about an *invariant scalar*.

3. Tensor of Rank 1

A *tensor of rank* 1 has three components t_1, t_2 and t_3. The transformation law (4.69) is now

$$\tilde{t}_{\mu} = \sum_{i=1}^{3} a_{\mu i} t_i \qquad (\mu = 1, 2, 3). \tag{4.70}$$

It is the transformation law for vectors, i.e., a vector is a tensor of rank 1.

4. Tensor of Rank 2

If $n = 2$, then the tensor $\boldsymbol{T}$ has nine components t_{ij}, which can be arranged in a matrix

$$\boldsymbol{T} = \mathbf{T} = \begin{pmatrix} t_{11} & t_{12} & t_{13} \\ t_{21} & t_{22} & t_{23} \\ t_{31} & t_{32} & t_{33} \end{pmatrix}. \tag{4.71a}$$

The transformation law (4.70) is now:

$$\tilde{t}_{\mu\nu} = \sum_{i=1}^{3}\sum_{j=1}^{3} a_{\mu i}a_{\nu j}t_{ij} \qquad (\mu, \nu = 1, 2, 3). \tag{4.71b}$$

So, a tensor of rank 2 can be represented as a matrix.

■ **A:** The moment of inertia Θ_g of a solid with respect to the line g, which goes through the origin and has direction vector $\vec{\mathbf{a}} = \underline{\mathbf{a}}^{\mathrm{T}}$, can be represented in the form

$$\Theta_g = \underline{\mathbf{a}}^{\mathrm{T}}\boldsymbol{\Theta}\underline{\mathbf{a}} \quad (4.72a) \qquad \text{with} \quad \boldsymbol{\Theta} = (\Theta_{ij}) = \begin{pmatrix} \Theta_x & -\Theta_{xy} & -\Theta_{xz} \\ -\Theta_{xy} & \Theta_y & -\Theta_{yz} \\ -\Theta_{xz} & -\Theta_{yz} & \Theta_z \end{pmatrix}, \quad (4.72b)$$

the so-called *inertia tensor*. Here Θ_x, Θ_y and Θ_z are the moments of inertia with respect to the coordinate axes, and Θ_{xy}, Θ_{xz} and Θ_{yz} are the *deviation moments* with respect to the coordinate axes.

■ **B:** The load-up conditions of an elastically deformed body can be given by the tension tensor

$$\boldsymbol{\sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{pmatrix}. \tag{4.73}$$

The elements σ_{ik} $(i, k = 1, 2, 3)$ are determined in the following way: At a point P of the elastic body a small plane surface element is chosen whose normal vector points to the direction of the x_1-axis of a right-angle Cartesian coordinate system. The power per surface unit on this element, depending on the material, is a vector with coordinates σ_{11}, σ_{12} and σ_{13}. The other components can be explained similarly.

5. Rules of Calculation

1. Elementary Algebraic Operations The multiplication of a tensor by a number, and addition and subtraction of tensors of the *same rank* are defined componentwise, similarly to the corresponding operations for vectors and matrices.

2. Tensor Product Suppose there are given a tensor $\boldsymbol{A}$ of rank m and a tensor $\boldsymbol{B}$ of rank n with components $a_{ij\cdots}$ and $b_{rs\cdots}$ respectively. Then the 3^{m+n} scalars

$$c_{ij\cdots rs\cdots} = a_{ij\cdots}b_{rs\cdots} \tag{4.74a}$$

give the components of a tensor $\boldsymbol{C}$ of rank $m+n$. It is denoted by $\boldsymbol{C} = \boldsymbol{AB}$ and it is called the *tensor product* of $\boldsymbol{A}$ and $\boldsymbol{B}$. The associative and distributive laws are valid:

$$(\boldsymbol{AB})\,\boldsymbol{C} = \boldsymbol{A}\,(\boldsymbol{BC}), \qquad \boldsymbol{A}\,(\boldsymbol{B} + \boldsymbol{C}) = \boldsymbol{AB} + \boldsymbol{AC}. \tag{4.74b}$$

3. Dyadic Product The product of two tensors of rank 1 $\boldsymbol{A} = (a_1, a_2, a_3)$ and $\boldsymbol{B} = (b_1, b_2, b_3)$ gives a tensor of rank 2 with the elements

$$c_{ij} = a_ib_j \qquad (i, j = 1, 2, 3), \tag{4.75a}$$

i.e., the tensor product results in the matrix

$$\begin{pmatrix} a_1b_1 & a_1b_2 & a_1b_3 \\ a_2b_1 & a_2b_2 & a_2b_3 \\ a_3b_1 & a_3b_2 & a_3b_3 \end{pmatrix}. \tag{4.75b}$$

This will be denoted as the *dyadic product* of the two vectors $\underline{\mathbf{A}}$ and $\underline{\mathbf{B}}$.

4. Contraction Setting two indices equal to each other in a tensor of rank m $(m \geq 2)$, and summing with respect to them, then one gets a tensor of rank $m-2$, which is called the *contraction* of the tensor.

■ The tensor $\boldsymbol{C}$ of rank 2 of (4.75a) with $c_{ij} = a_ib_j$, which is the tensor product of the vectors $\underline{\mathbf{A}} = (a_1, a_2, a_3)$ and $\underline{\mathbf{B}} = (b_1, b_2, b_3)$, can be contracted by the indices i and j,

$$a_ib_i = a_1b_1 + a_2b_2 + a_3b_3 \tag{4.76}$$

giving a scalar, which is a tensor of rank 0. This gives the scalar product of vectors $\underline{\mathbf{A}}$ and $\underline{\mathbf{B}}$.

4.3.3 Tensors with Special Properties

4.3.3.1 Tensors of Rank 2

1. Rules of Calculation

For tensors of rank 2 the same rules are valid as for matrices. In particular, every tensor $\boldsymbol{T}$ can be decomposed into the sum of a symmetric and a skew-symmetric tensor:

$$\boldsymbol{T} = \frac{1}{2}\left(\boldsymbol{T} + \boldsymbol{T}^{\mathrm{T}}\right) + \frac{1}{2}\left(\boldsymbol{T} - \boldsymbol{T}^{\mathrm{T}}\right). \tag{4.77a}$$

A tensor $\boldsymbol{T} = (t_{ij})$ is called *symmetric* if

$$t_{ij} = t_{ji} \qquad \text{for all } i \text{ and } j \tag{4.77b}$$

holds. In the case

$$t_{ij} = -t_{ji} \qquad \text{for all } i \text{ and } j \tag{4.77c}$$

it is called *skew- or antisymmetric*. Obviously the elements t_{11}, t_{22} and t_{33} of a skew-symmetric tensor are equal to zero. The notion of symmetry and antisymmetry can be extended for tensors of higher rank if refering to certain pairs of elements.

2. Transformation of Principal Axes

For a symmetric tensor $\boldsymbol{T}$, i.e., if $t_{\mu\nu} = t_{\nu\mu}$ holds, there is always an orthogonal transformation $\mathbf{D}$ such that after the transformation the tensor has a diagonal form:

$$\tilde{\boldsymbol{T}} = \begin{pmatrix} \tilde{t}_{11} & 0 & 0 \\ 0 & \tilde{t}_{22} & 0 \\ 0 & 0 & \tilde{t}_{33} \end{pmatrix}. \tag{4.78a}$$

The elements $\tilde{t}_{11}$, $\tilde{t}_{22}$ and $\tilde{t}_{33}$ are called the *eigenvalues of the tensor* $\boldsymbol{T}$. They are equal to the roots λ_1, λ_2 and λ_3 of the algebraic equation of third degree in λ:

$$\begin{vmatrix} t_{11}-\lambda & t_{12} & t_{13} \\ t_{21} & t_{22}-\lambda & t_{23} \\ t_{31} & t_{32} & t_{33}-\lambda \end{vmatrix} = 0. \tag{4.78b}$$

The column vectors $\underline{\mathbf{d}}_1$, $\underline{\mathbf{d}}_2$ and $\underline{\mathbf{d}}_3$ of the transformation matrix $\mathbf{D}$ are called the *eigenvectors* corresponding to the eigenvalues, and they satisfy the equations

$$\boldsymbol{T}\underline{\mathbf{d}}_\nu = \lambda_\nu \underline{\mathbf{d}}_\nu \qquad (\nu = 1, 2, 3). \tag{4.78c}$$

Their directions are called the *directions of the principal axes*, and the transformation $\boldsymbol{T}$ to diagonal form is called the *transformation of the principal axes*.

4.3.3.2 Invariant Tensors

1. Definition

A Cartesian tensor is called *invariant* if its components are the same in all Cartesian coordinate systems. Physical quantities such as scalars and vectors, which are special tensors, do not depend on the coordinate system in which they are determined; they must not change their value either under translation of the origin or rotation of a coordinate system K. One talks about *translation invariance* and about *rotation invariance* or in general about *transformation invariance*.

2. Generalized Kronecker Delta or Delta Tensor

If the elements t_{ij} of a tensor of rank 2 are the Kronecker symbols, i.e.,

$$t_{ij} = \delta_{ij} = \begin{cases} 1 & \text{for } i = j, \\ 0 & \text{otherwise,} \end{cases} \tag{4.79a}$$

then from the transformation law (4.71b) in the case of a rotation of the coordinate system considering (4.68c) follows

$$\tilde{t}_{\mu\nu} = d_{\mu i} d_{\nu j} = \delta_{\mu\nu}\,, \tag{4.79b}$$

i.e., the elements are *rotation invariant.* Putting them into a coordinate system so that they are independent of the choice of the origin, i.e., they will be *translation invariant*, then the numbers δ_{ij} form a tensor of rank 2, the so-called *generalized Kronecker delta* or *delta tensor*.

3. Alternating Tensor

If $\vec{e}_i$, $\vec{e}_j$ and $\vec{e}_k$ are unit vectors in the directions of the axes of a right-angle coordinate system, then for the mixed product (see 3.5.1.6, **2.**, p. 185) holds

$$\epsilon_{ijk} = \vec{e}_i\,(\vec{e}_j \times \vec{e}_k) = \begin{cases} 1, & \text{if } i, j, k \text{ cyclic (right-hand rule)}, \\ -1, & \text{if } i, j, k \text{ anticyclic}, \\ 0, & \text{otherwise}. \end{cases} \tag{4.80a}$$

Altogether there are $3^3 = 27$ elements, which are the elements of a tensor of rank 3. In the case of a rotation of the coordinate system from the transformation law (4.69) it follows that

$$\tilde{t}_{\mu\nu\rho} = d_{\mu i} d_{\nu j} d_{\rho k} \epsilon_{ijk} = \begin{vmatrix} d_{\mu 1} & d_{\nu 1} & d_{\rho 1} \\ d_{\mu 2} & d_{\nu 2} & d_{\rho 2} \\ d_{\mu 3} & d_{\nu 3} & d_{\rho 3} \end{vmatrix} = \epsilon_{\mu\nu\rho}, \tag{4.80b}$$

i.e., the elements are *rotation invariant.* Putting them into a coordinate system so that they are independent of the choice of the origin, i.e., they are *translation invariant*, then the numbers ϵ_{ijk} form a tensor of rank 3, the so-called *alternating tensor.*

4. Tensor Invariants

There must not be a confusion between *tensor invariants* and invariant tensors. Tensor invariants are functions of the components of tensors whose forms and values do not change during the rotation of the coordinate system.

■ **A:** If for instance the tensor $\boldsymbol{T} = (t_{ij})$ is transformed in $\tilde{\boldsymbol{T}} = (\tilde{t}_{ij})$ by a rotation, then the *trace* (*spur*) of it does not change:

$$\mathrm{Tr}(\boldsymbol{T}) = t_{11} + t_{22} + t_{33} = \tilde{t}_{11} + \tilde{t}_{22} + \tilde{t}_{33}\,. \tag{4.81}$$

The trace of the tensor $\boldsymbol{T}$ is equal to the sum of the eigenvalues (see 4.1.2, **4.**, p. 270).

■ **B:** For the determinant of the tensor $\boldsymbol{T} = (t_{ij})$

$$\begin{vmatrix} t_{11} & t_{12} & t_{13} \\ t_{21} & t_{22} & t_{23} \\ t_{31} & t_{32} & t_{33} \end{vmatrix} = \begin{vmatrix} \tilde{t}_{11} & \tilde{t}_{12} & \tilde{t}_{13} \\ \tilde{t}_{21} & \tilde{t}_{22} & \tilde{t}_{23} \\ \tilde{t}_{31} & \tilde{t}_{32} & \tilde{t}_{33} \end{vmatrix} \tag{4.82}$$

is valid. The determinant of the tensor is equal to the product of the eigenvalues.

4.3.4 Tensors in Curvilinear Coordinate Systems

4.3.4.1 Covariant and Contravariant Basis Vectors

1. Covariant Basis

By the help of the variable position vector are introduced the general *curvilinear coordinates* u, v, w:

$$\vec{r} = \vec{r}\,(u, v, w) = x(u, v, w)\vec{e}_x + y(u, v, w)\vec{e}_y + z(u, v, w)\vec{e}_z. \tag{4.83a}$$

The *coordinate surfaces* corresponding to this system can be got by fixing the independent variables u, v, w in $\vec{r}\,(u, v, w)$, one at a time. There are three coordinate surfaces passing through every point of the considered region of space, and any two of them intersect each other in a *coordinate line*, and of course these curves pass through the considered point, too. The three vectors

$$\frac{\partial \vec{r}}{\partial u}, \quad \frac{\partial \vec{r}}{\partial v}, \quad \frac{\partial \vec{r}}{\partial w} \tag{4.83b}$$

point along the directions of the coordinate lines in the considered point. They form the *covariant basis* of the curvilinear coordinate system.

2. Contravariant Basis

The three vectors

$$\frac{1}{D}\left(\frac{\partial\vec r}{\partial v}\times\frac{\partial\vec r}{\partial w}\right),\quad \frac{1}{D}\left(\frac{\partial\vec r}{\partial w}\times\frac{\partial\vec r}{\partial u}\right),\quad \frac{1}{D}\left(\frac{\partial\vec r}{\partial u}\times\frac{\partial\vec r}{\partial v}\right) \tag{4.84a}$$

with the functional determinant (Jacobian determinant see 2.18.2.6,**3.**, p. 123)

$$D=\frac{D(x,y,z)}{D(u,v,w)}=\begin{vmatrix} x_u & x_v & x_w\\ y_u & y_v & y_w\\ z_u & z_v & z_w\end{vmatrix} \tag{4.84b}$$

are always perpendicular to the coordinate surfaces at the considered surface element and they form the so-called *contravariant basis* of the curvilinear coordinate system.

Remark: In the case of orthogonal curvilinear coordinates, i.e., if

$$\frac{D(x,y,z)}{D(u,v,w)}=\begin{vmatrix} t_{11} & t_{12} & t_{13}\\ t_{21} & t_{22} & t_{23}\\ t_{31} & t_{32} & t_{33}\end{vmatrix}=\begin{vmatrix} \tilde t_{11} & \tilde t_{12} & \tilde t_{13}\\ \tilde t_{21} & \tilde t_{22} & \tilde t_{23}\\ \tilde t_{31} & \tilde t_{32} & \tilde t_{33}\end{vmatrix}\ \frac{\partial\vec r}{\partial u}\cdot\frac{\partial\vec r}{\partial v}=0,\quad \frac{\partial\vec r}{\partial u}\cdot\frac{\partial\vec r}{\partial w}=0,\quad \frac{\partial\vec r}{\partial v}\cdot\frac{\partial\vec r}{\partial w}=0, \tag{4.85}$$

then the directions of the covariant and contravariant basis are coincident.

4.3.4.2 Covariant and Contravariant Coordinates of Tensors of Rank 1

In order to be able to apply the summation convention of Einstein the following notation is introduced for the covariant and contravariant basis:

$$\begin{aligned}&\frac{\partial\vec r}{\partial u}=\vec g_1,\quad \frac{\partial\vec r}{\partial v}=\vec g_2,\quad \frac{\partial\vec r}{\partial w}=\vec g_3\quad\text{and}\\ &\frac{1}{D}\left(\frac{\partial\vec r}{\partial v}\times\frac{\partial\vec r}{\partial w}\right)=\vec g^{\,1},\quad \frac{1}{D}\left(\frac{\partial\vec r}{\partial w}\times\frac{\partial\vec r}{\partial u}\right)=\vec g^{\,2},\quad \frac{1}{D}\left(\frac{\partial\vec r}{\partial u}\times\frac{\partial\vec r}{\partial v}\right)=\vec g^{\,3}.\end{aligned} \tag{4.86}$$

Then the following representations hold for $\vec v$:

$$\vec v=V^1\vec g_1+V^2\vec g_2+V^3\vec g_3=V^k\vec g_k\quad\text{or}\quad \vec v=V_1\vec g^{\,1}+V_2\vec g^{\,2}+V_3\vec g^{\,3}. \tag{4.87}$$

The components V^k are the contravariant coordinates, the components V_k are the covariant coordinates of the vector $\vec v$. For these coordinates the equalities

$$V^k=g^{kl}V_l\qquad\text{and}\qquad V_k=g_{kl}V^l \tag{4.88a}$$

are valid, where

$$g_{kl}=g_{lk}=\vec g_k\cdot\vec g_l\qquad\text{and}\qquad g^{kl}=g^{lk}=\vec g^{\,k}\cdot\vec g^{\,l} \tag{4.88b}$$

respectively. Furthermore using the Kronecker symbol the equality

$$\vec g_k\cdot\vec g^{\,l}=\delta_{kl}, \tag{4.89a}$$

holds, and consequently

$$g^{kl}g_{lm}=\delta_{km}. \tag{4.89b}$$

The transition from V^k to V_k or from V_k to V^k according to (4.88b) is described by raising or lowering the indices by *oversliding.*

Remark: In Cartesian coordinate systems covariant and contravariant coordinates are equal to each

other.

4.3.4.3 Covariant, Contravariant and Mixed Coordinates of Tensors of Rank 2

1. Coordinate Transformation

In a Cartesian coordinate system with basis vectors $\vec{e}_1$, $\vec{e}_2$ and $\vec{e}_3$ a tensor $\boldsymbol{T}$ of rank 2 can be represented as a matrix

$$\mathbf{T} = \begin{pmatrix} t_{11} & t_{12} & t_{13} \\ t_{21} & t_{22} & t_{23} \\ t_{31} & t_{32} & t_{33} \end{pmatrix}. \tag{4.90}$$

To introduce curvilinear coordinates u_1, u_2, u_3 the following vector is used:

$$\vec{r} = x_1(u_1, u_2, u_3)\vec{e}_1 + x_2(u_1, u_2, u_3)\vec{e}_2 + x_3(u_1, u_2, u_3)\vec{e}_3\,. \tag{4.91}$$

The new basis is denoted by the vectors $\vec{g}_1$, $\vec{g}_2$ and $\vec{g}_3$. Now it holds:

$$\vec{g}_l = \frac{\partial \vec{r}}{\partial u_l} = \frac{\partial x_1}{\partial u_l}\vec{e}_1 + \frac{\partial x_2}{\partial u_l}\vec{e}_2 + \frac{\partial x_3}{\partial u_l}\vec{e}_3 = \frac{\partial x_k}{\partial u_l}\vec{e}_k. \tag{4.92}$$

Substituting $\vec{e}_l = \vec{g}^{\,l}$, then follows $\vec{g}_l$ and $\vec{g}^{\,l}$ as covariant and contravariant basis vectors.

2. Linear Vector Function

In a fixed coordinate system with the tensor $\boldsymbol{T}$ given as in (4.90) by the equality

$$\vec{w} = \boldsymbol{T}\vec{v} \tag{4.93a}$$

the following vector representations

$$\vec{v} = V_k\vec{g}^{\,k} = V^k\vec{g}_k, \quad \vec{w} = W_k\vec{g}^{\,k} = W^k\vec{g}_k \tag{4.93b}$$

define a linear relation between the vectors $\vec{v}$ and $\vec{w}$. So (4.93a) is to be considered as a *linear vector function.*

3. Mixed Coordinates

Changing the coordinate system, the equality (4.93a) will have the form

$$\tilde{\vec{w}} = \tilde{\boldsymbol{T}}\tilde{\vec{v}}\,. \tag{4.94a}$$

The relation between the components of $\boldsymbol{T}$ and $\tilde{\boldsymbol{T}}$ is the following:

$$\tilde{t}_{kl} = \frac{\partial u_k}{\partial x_m}\frac{\partial x_n}{\partial u_l}t_{mn}\,. \tag{4.94b}$$

Introducing the notation

$$\tilde{t}_{kl} = T^k_{\cdot l} \tag{4.94c}$$

one talks about *mixed coordinates* of the tensor; k contravariant index, l covariant index. For the components of vectors $\vec{v}$ and $\vec{w}$ holds

$$W^k = T^k_{\cdot l}V^l\,. \tag{4.94d}$$

If the covariant basis $\vec{g}_k$ is replaced by the contravariant basis $\vec{g}^{\,k}$, then one gets similarly to (4.94b) and (4.94c)

$$T_k^{\cdot\,l} = \frac{\partial x_m}{\partial u_k}\frac{\partial u_l}{\partial x_n}t_{mn}\,, \tag{4.95a}$$

and (4.94d) is transformed into

$$W_k = T_k^{\cdot\,l}V_l. \tag{4.95b}$$

For the mixed coordinates $T_k^{\cdot\,l}$ and $T^k_{\cdot l}$ holds the formula

$$T^k_{\cdot l} = g^{km}g_{ln}T_m^{\cdot\,n}\,. \tag{4.95c}$$

4. Pure Covariant and Pure Contravariant Coordinates

Substituting in (4.95b) for V_l the relation $V_l = g_{lm}V^m$, then one gets

$$W_k = T_k^{\cdot l} g_{lm} V^m = T_{km} V^m \,, \tag{4.96a}$$

also considering that

$$T_k^{\cdot l} g_{lm} = T_{km} \,. \tag{4.96b}$$

The T_{km} are called the covariant coordinates of the tensor $\boldsymbol{T}$, because both indices are covariant. Similarly one gets the contravariant coordinates

$$T_l^{km} = g^{ml} T^k_{\cdot l} \,. \tag{4.97}$$

The explicit forms are:

$$T_{kl} = \frac{\partial x_m}{\partial u_k}\frac{\partial x_n}{\partial u_l} t_{mn} \,, \tag{4.98a}$$

$$T^{kl} = \frac{\partial u_k}{\partial x_m}\frac{\partial u_l}{\partial x_n} t_{mn} \,. \tag{4.98b}$$

4.3.4.4 Rules of Calculation

In addition to the rules described on 4.3.2, **5.**, p. 283, the following rules of calculations are valid:

1. Addition, Subtraction Tensors of the same rank whose corresponding indices are both covariant or contravariant can be added or subtracted elementwise, and the result is a tensor of the same rank.

2. Multiplication The multiplication of the coordinates of a tensor of rank n by the coordinates of a tensor of rank m results in a tensor of rank $m + n$.

3. Contraction If making the indices of a covariant and a contravariant coordinates of a tensor of rank n $(n \geq 2)$ equal, then can be used the Einstein summation convention for this index, and one gets a tensor of rank $n - 2$. This operation is called *contraction*.

4. Oversliding *Oversliding* of two tensors is the following operation: Multiply both, then making a contraction so that the indices by which the contraction is made belong to different factors.

5. Symmetry A tensor is called symmetric with respect to two covariant or two contravariant standing indices if when exchanging them the tensor does not change.

6. Skew-Symmety A tensor is called skew-symmetric with respect to two covariant or two contravariant standing indices if when exchanging them the tensor is multiplied by -1.

■ The alternating tensor (see 4.3.3.2, **3.**,p. 284) is skew-symmetric with respect to two arbitrary covariant or contravariant indices.

4.3.5 Pseudotensors

The reflection of a tensor plays a special role in physics. Because of their different behavior with respect to reflection *polar* and *axial vectors* are distinguished (see 3.5.1.1, **2.**, p. 181), although mathematically they can be handled in the same way. Axial and polar vectors differ from each other in their determination, because axial vectors can be represented by an orientation in addition to length and direction. Axial vectors are also called *pseudovectors*. Since vectors can be considered as tensors, the general notion of pseudotensors is introduced.

4.3.5.1 Symmetry with Respect to the Origin

1. Behavior of Tensors under Space Inversion

1. Notion of Space Inversion The *reflection of the position coordinates* of points in space with respect to the origin is called *space inversion* or *coordinate inversion*. In a three-dimensional Cartesian coordinate system space inversion means the change of the sign of the coordinates:

$$(x, y, z) \to (-x, -y, -z). \tag{4.99}$$

By this a right-hand coordinate system becomes a left-hand system. Similar rules are valid for other coordinate systems. In the spherical coordinate system holds:

$$(r, \vartheta, \varphi) \to (-r, \pi - \vartheta, \varphi + \pi). \tag{4.100}$$

Under this type of reflection the length of the vectors and the angles between them do not change. The transition can be given by a linear transformation.

2. Transformation Matrix According to (4.66), the transformation matrix $\mathbf{A} = (a_{\mu\nu})$ of a linear transformation of three-dimensional space has the following properties in the case of space inversion:

$$a_{\mu\nu} = -\delta_{\mu\nu}, \quad \det \mathbf{A} = -1. \tag{4.101a}$$

For the components of a tensor of rank n (4.69)

$$\tilde{t}_{\mu\nu\cdots\rho} = (-1)^n t_{\mu\nu\cdots\rho} \tag{4.101b}$$

holds. That is: In the case of point symmetry with respect to the origin a tensor of rank 0 remains a scalar, unchanged; a tensor of rank 1 remains a vector with a change of sign; a tensor of rank 2 remains unchanged, etc.

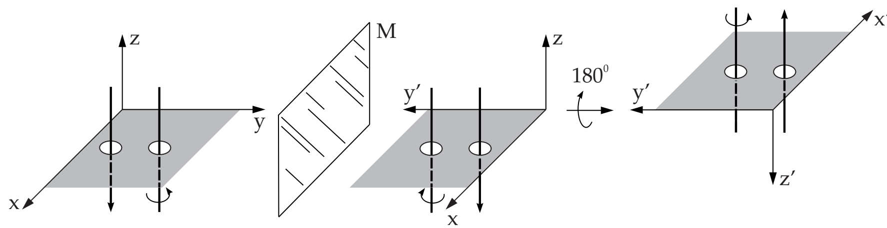


Figure 4.3

2. Geometric Representation

The inversion of space in a three-dimensional Cartesian coordinate system can be realized in two steps **(Fig.4.3)**:

1. By reflection with respect to the coordinate plane, for instance the x, z plane, the coordinate system x, y, z turns into the coordinate system $x, -y, z$. A right-hand system becomes a left-hand system (see 3.5.3.1,**2.**, p. 209).

2. By a rotation of the system x, y, z around the y-axis by 180° we have the complete coordinate system x, y, z reflected with respect to the origin. This coordinate system stays left-handed, as it was after the first step.

Conclusion: Space inversion changes the orientation of a polar vector by 180°, while an axial vector keeps its orientation.

4.3.5.2 Introduction to the Notion of Pseudotensors

1. Vector Product under Space Inversion Under space inversion two polar vectors $\underline{\mathbf{a}}$ and $\underline{\mathbf{b}}$ are transformed into the vectors $-\underline{\mathbf{a}}$ and $-\underline{\mathbf{b}}$, i.e., their components satisfy the transformation formula (4.101b) for tensors of rank 1. However, if considering the vector product $\underline{\mathbf{c}} = \underline{\mathbf{a}} \times \underline{\mathbf{b}}$ as an example of an axial vector, then one gets $\underline{\mathbf{c}} = \underline{\mathbf{c}}$ under reflection with respect to the origin. This is a violation of the transformation formula (4.101a) for tensors of rank 1. Therefore the axial vector $\underline{\mathbf{c}}$ is called a *pseudovector* or generally a *pseudotensor*.

■ The vector products $\vec{r} \times \vec{v}$, $\vec{r} \times \vec{F}$, $\nabla \times \vec{v} = \operatorname{rot} \vec{v}$ with the position vector $\vec{r}$, the speed vector $\vec{v}$, the power vector $\vec{F}$ and the nabla operator ∇ are examples of axial vectors, which have "false" behavior under reflection.

2. Scalar Product under Space Inversion If using space inversion for a scalar product of a polar and an axial vector, then again there is a case of violation of the transformation formula (4.101b) for tensors of rank 1. Because the result of a scalar product is a scalar, and a scalar should be the

same in every coordinate system, here it is a very special scalar, which is called a *pseudoscalar*. It has the property that it changes its sign under space inversion. Pseudoscalars do not have the *rotation invariance property* of scalars.

■ The scalar product of the polar vectors $\vec{r}$ (position vector) and $\vec{v}$ (speed vector) by the axial vector $\vec{\omega}$ (angular velocity vector) results in the scalars $\vec{r}\cdot\vec{\omega}$ and $\vec{v}\cdot\vec{\omega}$, which have the "false" behavior under reflection, so they are pseudoscalars.

3. Mixed Product under Space Inversion The mixed product $(\underline{\mathbf{a}}\times\underline{\mathbf{b}})\cdot\underline{\mathbf{c}}$ (see 3.5.1.6, **2.**, p. 185) of polar vectors $\underline{\mathbf{a}}$, $\underline{\mathbf{b}}$, and $\underline{\mathbf{c}}$ is a *pseudoscalar* according to (2.), because the factor $(\underline{\mathbf{a}}\times\underline{\mathbf{b}})$ is an axial vector. The sign of the mixed product changes under space inversion.

4. Pseudovector and Skew-Symmetric Tensor of Rank 2 The tensor product of axial vectors $\underline{\mathbf{a}} = (a_1, a_2, a_3)^{\mathrm{T}}$ and $\underline{\mathbf{b}} = (b_1, b_2, b_3)^{\mathrm{T}}$ results in a tensor of rank 2 with components $t_{ij} = a_i b_j \quad (i, j = 1, 2, 3)$ according to (4.74a). Since every tensor of rank 2 can be decomposed into a sum of a symmetric and a skew-symmetric tensor of rank 2, according to (4.81)

$$t_{ij} = \frac{1}{2}(a_i b_j + a_j b_i) + \frac{1}{2}(a_i b_j - a_j b_i) \quad (i, j = 1, 2, 3) \tag{4.102}$$

holds. The skew-symmetric part of (4.102) contains exactly the components of the vector product $(\underline{\mathbf{a}}\times\underline{\mathbf{b}})$ multiplied by $\frac{1}{2}$, so the axial vector $\underline{\mathbf{c}} = (\underline{\mathbf{a}}\times\underline{\mathbf{b}})$ with components c_1, c_2, c_3 can be considered as a skew-symmetric tensor of rank 2

$$\boldsymbol{C} = \underline{\mathbf{c}} = \begin{pmatrix} 0 & c_{12} & c_{13} \\ -c_{12} & 0 & c_{23} \\ -c_{13} & -c_{23} & 0 \end{pmatrix} \quad \text{(4.103a)} \quad \text{where} \quad \begin{aligned} c_{23} &= a_2 b_3 - a_3 b_2 = c_1, \\ c_{31} &= a_3 b_1 - a_1 b_3 = c_2, \\ c_{12} &= a_1 b_2 - a_2 b_1 = c_3, \end{aligned} \quad \text{(4.103b)}$$

whose components satisfy the transformation formula (4.101b) for tensors of rank 2.
Consequently every axial vector (pseudovector or pseudotensor of rank 1) $\underline{\mathbf{c}} = (c_1, c_2, c_3)^{\mathrm{T}}$ can be considered as a skew-symmetric tensor $\boldsymbol{C}$ of rank 2:

$$\boldsymbol{C} = \underline{\mathbf{c}} = \begin{pmatrix} 0 & c_3 & -c_2 \\ -c_3 & 0 & c_1 \\ c_2 & -c_1 & 0 \end{pmatrix}. \tag{4.104}$$

5. Pseudotensors of Rank *n* The generalization of the notion of pseudoscalar and pseudovector is a pseudotensor of rank n. It has the same property under rotation as a tensor of rank n (rotation matrix $\mathbf{D}$ with $\det \mathbf{D} = 1$) but it has a (-1) factor under reflection through the origin. Examples of pseudotensors of higher rank can be found in the literature, e.g., [4.1].

4.4 Quaternions and Applications

Quaternions were defined by Sir William Rowan Hamilton in 1843. The basic question which resulted the discovering of quaternions was that how could division of vectors in the three dimensional Euclidian space be defined. It is possible by embedding them into $\mathbb{R}^4$, and introducing the quaternion multiplication, what leads to the *division ring* of quaternions.

Quaternions, like complex numbers, both are special cases of a Clifford-algebra of order n, with 2^n generalized numbers as basis (see [4.5], [22.22]):

$$A = \sum_{l=1}^{2^n} \mathbf{i}_l\, a_l \quad (\mathbf{i}_l \text{ hyper-complex elements, } a_l \text{ complex numbers}). \tag{4.105a}$$

The following special cases have practical importance:

$n = 1$: 2-dimensional complex numbers with

$$\mathbf{i}_1 = 1, \mathbf{i}_2 = \mathrm{i} \ \text{(i imaginary unit)}, \quad a_1, a_2 \ \text{(real numbers)}. \tag{4.105b}$$

$n = 2$: Quaternions as 4-dimensional numbers with hyper-complex elements

$$\mathbf{i}_1 = 1,\ \mathbf{i}_2 = \mathbf{i},\ \mathbf{i}_3 = \mathbf{j},\ \mathbf{i}_4 = \mathbf{k} \ \text{(hyper-complex elements)}, \quad a_1, a_2, a_3, a_4 \ \text{(real numbers)} \tag{4.105c}$$

and the rules of multiplication (4.107). In physics the PAULI's spin matrices and spinors are represented as quaternions.

$n = 3$: Biquaternions (s. 4.4.3.6, **1.**, p. 306)

$n = 4$: Clifford-numbers are known in physics as DIRAC-matrices.

Quaternions are used most often to describe rotations. The advantages of the quaternions are:
• the rotation is performed directly around the required axis,
• the gimbal-lock problem does not occur. Gimbal is a pivoted support allowing the rotation around a single axis (e.g. gyrocompass), and gimbal lock means that the axes of two of the three gimbals are driven into parallel configuration.

The disadvantage of quaternions is that only rotations can be described with them. To represent translations, scaling or projections matrices are needed. This disadvantage can be overcome by biquaternions, by which every motion of rigid bodies can be represented.

Quaternion are used in computer-graphics, satellite-navigation, in vector analysis, in physics and in mechanics.

4.4.1 Quaternions

4.4.1.1 Definition and Representation

1. Imaginary Units

Quaternions are generalized complex numbers in the form

$$w + \mathbf{i}x + \mathbf{j}y + \mathbf{k}z, \tag{4.106}$$

where w, x, y, z are real numbers, and the generalized imaginary units are $\mathbf{i}$, $\mathbf{j}$,$\mathbf{k}$, which satisfy the following rules of multiplication:

$$\mathbf{i}^2 = \mathbf{j}^2 = \mathbf{k}^2 = -1, \qquad \mathbf{ij} = \mathbf{k} = -\mathbf{ji}, \quad \mathbf{jk} = \mathbf{i} = -\mathbf{kj}, \quad \mathbf{ki} = \mathbf{j} = -\mathbf{ik}. \tag{4.107}$$

	i	**j**	**k**
i	-1	**k**	$-$**j**
j	$-$**k**	-1	**i**
k	**j**	$-$**i**	-1

Multiplication table

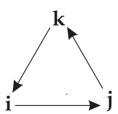


Figure 4.4

The multiplication of the generalized units is shown in the accompanying table of multiplication. Alternatively the multiplication rules can be represented by the cycle shown in **Fig.4.4**. Multiplication in the direction of an arrow results in a positive sign, opposite to the arrow direction results in a negative sign.

Consequently, the multiplication is not commutative but associative. The four-dimensional Euclidian vector space $\mathbb{R}^4$ provided with quaternion-multiplication is denoted by $\mathbb{H}$ in honour of R.W. HAMILTON. Quaternions form an algebra, namely the *division ring of quaternions*. .

2. Representation of Quaternions

There are different representation of quaternions:
• as hyper-complex numbers $q = w + \mathbf{i}x + \mathbf{j}y + \mathbf{k}z = q_0 + \underline{\mathbf{q}}$ with scalar part $q_0 = \mathrm{Sc}q$ and vector part $\underline{\mathbf{q}} = \mathrm{Vec}q$,
• as four dimensional vector $q = (w, x, y, z)^T = (q_0, \underline{\mathbf{q}})^T$ consisting of the number $w \in \mathbb{R}$ and the vector $(x, y, z)^T \in \mathbb{R}^3$,
• in trigonometric form $q = r(\cos\varphi + \underline{\mathbf{n}}_{\mathbf{q}} \sin\varphi)$, where $r = |q| = \sqrt{w^2 + x^2 + y^2 + z^2}$ is the length of the four dimensional vector in $\mathbb{R}^4$, and $\cos\varphi = \dfrac{w}{|q|}$, and $\underline{\mathbf{n}}_{\mathbf{q}} = \dfrac{1}{|(x, y, z)^T|}(x, y, z)^T$. $\underline{\mathbf{n}}_{\mathbf{q}}$ is a unit vector in $\mathbb{R}^3$, depending on q.

Remark: The multiplication rules for quaternions differ from the usually introduced rules in $\mathbb{R}^3$ and

$\mathbb{R}^4$ (see (4.109b), (4.114), (4.115)).

3. Relation between Hyper-complex Number and Trigonometric Form

$$q = q_0 + \underline{\mathbf{q}} = |q|\left(\frac{q_0}{|q|} + \frac{\underline{\mathbf{q}}}{|q|}\right) = |q|\left(\frac{q_0}{|q|} + \frac{\underline{\mathbf{q}}}{|\underline{\mathbf{q}}|}\frac{|\underline{\mathbf{q}}|}{|q|}\right) = r(\cos\varphi + \underline{\mathbf{n}}_{\underline{\mathbf{q}}}\sin\varphi)\,, \tag{4.108a}$$

if $|\underline{\mathbf{q}}| \neq 0$. If $|\underline{\mathbf{q}}| = 0$, then there is

$$q = q_0 = |q_0|\frac{q_0}{|q_0|} = \begin{cases} |q_0| = |q_0|\cos 0 & \text{for } q_0 > 0, \\ |q_0|(-1) = |q_0|\cos\pi & \text{for } q_0 < 0, \end{cases} \tag{4.108b}$$

if $q_0 \neq 0$.

4. Pure Quaternions

A pure quaternion has a zero scalar part: $q_0 = 0$. The set of pure quaternions is denoted by $\mathbb{H}_0$. It is often useful to identify the pure quaternions $\underline{\mathbf{q}}$ with the geometric vector $\vec{\mathbf{q}} \in \mathbb{R}^3$, i.e.

$$q = q_0 + \begin{cases} \underline{\mathbf{q}}, & \text{if } \underline{\mathbf{q}} \text{ represents a pure quaternion,} \\ \vec{\mathbf{q}}, & \text{if } \underline{\mathbf{q}} \text{ is interpreted as a geometric vector.} \end{cases} \tag{4.109a}$$

The *multiplication rule* for $\underline{\mathbf{p}}$, $\underline{\mathbf{q}} \in \mathbb{H}_0$ is

$$\underline{\mathbf{p}}\,\underline{\mathbf{q}} = -\vec{\mathbf{p}}\cdot\vec{\mathbf{q}} + \vec{\mathbf{p}}\times\vec{\mathbf{q}}\,, \tag{4.109b}$$

where $\cdot$ and $\times$ denote the dot-product and cross-product in $\mathbb{R}^3$, respectively. The result of (4.109b) is to interpret as a quaternion.

■ Let $\nabla = \frac{\partial}{\partial x}\vec{\mathbf{i}} + \frac{\partial}{\partial y}\vec{\mathbf{j}} + \frac{\partial}{\partial z}\vec{\mathbf{k}}$ be the nabla-operator (see 13.2.6.1, p. 715), and let $\vec{\mathbf{v}} = v_1(x,y,z)\vec{\mathbf{i}} + v_2(x,y,z)\vec{\mathbf{j}} + v_3(x,y,z)\vec{\mathbf{k}}$ be a vector-field. Here $\vec{\mathbf{i}}, \vec{\mathbf{j}}, \vec{\mathbf{k}}$ are unit-vectors being parallel to the coordinate axes in a Cartesian coordinate system. If ∇ and $\vec{\mathbf{v}}$ are interpreted as pure quaternions then according to (4.107) their product is:

$$\nabla\vec{\mathbf{v}} = -\frac{\partial v_1}{\partial x} - \frac{\partial v_2}{\partial y} - \frac{\partial v_3}{\partial z} + \vec{\mathbf{i}}\left(\frac{\partial v_3}{\partial y} - \frac{\partial v_2}{\partial z}\right) + \vec{\mathbf{j}}\left(\frac{\partial v_1}{\partial z} - \frac{\partial v_3}{\partial x}\right) + \vec{\mathbf{k}}\left(\frac{\partial v_2}{\partial x} - \frac{\partial v_1}{\partial y}\right).$$

This quaternion can be written in vector interpretation:

$$\nabla\vec{\mathbf{v}} = -\operatorname{div}\vec{\mathbf{v}} + \operatorname{rot}\vec{\mathbf{v}}\,,$$

but the result should be considered as a quaternion.

5. Unit Quaternions

A quaternion q is a unit quaternion if $|q| = 1$. The set of unit quaternions is denoted by $\mathbb{H}_1$. $\mathbb{H}_1$ is a so-called multiplicative Lie-group. The set $\mathbb{H}_1$ can be identified with the three dimensional sphere $S^3 = \{\underline{\mathbf{x}} \in \mathbb{R}^4 : |\underline{\mathbf{x}}| = 1\}$.

4.4.1.2 Matrix Representation of Quaternions

1. Real Matrices

If the number 1 is identified with the identity matrix

$$1 \triangleq \begin{pmatrix} 1&0&0&0\\ 0&1&0&0\\ 0&0&1&0\\ 0&0&0&1 \end{pmatrix} \tag{4.110a}$$

furthermore

$$\mathbf{i} \triangleq \begin{pmatrix} 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \end{pmatrix}, \quad \mathbf{j} \triangleq \begin{pmatrix} 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}, \quad \mathbf{k} \triangleq \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \end{pmatrix}, \tag{4.110b}$$

then a quaternion $q = w + \mathbf{i}x + \mathbf{j}y + \mathbf{k}z$ can be represented as a matrix

$$q \triangleq \begin{pmatrix} w & -x & -y & z \\ x & w & -z & -y \\ y & z & w & x \\ -z & y & -x & w \end{pmatrix}. \tag{4.110c}$$

2. Complex Matrices

Quaternions can be represented by complex matrices:

$$\mathbf{i} \triangleq \begin{pmatrix} 0 & -\mathrm{i} \\ -\mathrm{i} & 0 \end{pmatrix}, \quad \mathbf{j} \triangleq \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \quad \mathbf{k} \triangleq \begin{pmatrix} -\mathrm{i} & 0 \\ 0 & \mathrm{i} \end{pmatrix}. \tag{4.111a}$$

So

$$q = w + \mathbf{i}x + \mathbf{j}y + \mathbf{k}z \triangleq \begin{pmatrix} w - \mathrm{i}z & -\mathrm{i}x - y \\ -\mathrm{i}x + y & w + \mathrm{i}z \end{pmatrix}. \tag{4.111b}$$

Remarks:
1. On the right hand side of equations (4.111a,b) i represents the imaginary unit of complex numbers.
2. Matrix representation of quaternions is not unique, i.e. it is possible to give representations different from the ones in (4.110b,c) and (4.111a,b).

3. Conjugate and inverse Element

The conjugate of quaternion $q = w + \mathbf{i}x + \mathbf{j}y + \mathbf{k}z$ is the quaternion

$$\bar{q} = w - \mathbf{i}x - \mathbf{j}y - \mathbf{k}z\,. \tag{4.112a}$$

Obviously

$$|q|^2 = q\,\bar{q} = \bar{q}\,q = w^2 + x^2 + y^2 + z^2\,. \tag{4.112b}$$

Consequently every quaternion $q \in \mathbb{H} \setminus \{0\}$ has an inverse element

$$q^{-1} = \frac{\bar{q}}{|q|^2}\,. \tag{4.112c}$$

4.4.1.3 Calculation Rules

1. Addition and Subtraction

Addition and subtraction of two or more quaternions are defined as

$$\begin{aligned} &q_1 + q_2 - q_3 + \ldots \\ &= (w_1 + \mathbf{i}x_1 + \mathbf{j}y_1 + \mathbf{k}z_1) + (w_2 + \mathbf{i}x_2 + \mathbf{j}y_2 + \mathbf{k}z_2) - (w_3 + \mathbf{i}x_3 + \mathbf{j}y_3 + \mathbf{k}z_3) + \ldots \\ &= (w_1 + w_2 - w_3 + \ldots) + \mathbf{i}(x_1 + x_2 - x_3 + \ldots) + \mathbf{j}(y_1 + y_2 - y_3 + \ldots) \\ &\quad + \mathbf{k}(z_1 + z_2 - z_3 + \ldots)\,. \end{aligned} \tag{4.113}$$

Quaternions are added and subtracted as vectors in $\mathbb{R}^4$, or as matrices.

2. Multiplication

The multiplication is associative, so

$$\begin{aligned} q_1 q_2 = &\ (w_1 + \mathbf{i}x_1 + \mathbf{j}y_1 + \mathbf{k}z_1)(w_2 + \mathbf{i}x_2 + \mathbf{j}y_2 + \mathbf{k}z_2) \\ &= (w_1 w_2 - x_1 x_2 - y_1 y_2 - z_1 z_2) + \mathbf{i}(w_1 x_2 + w_2 x_1 + y_1 z_2 - z_1 y_2) + \\ &+ \mathbf{j}(w_1 y_2 + w_2 y_1 + z_1 x_2 - z_2 x_1) + \mathbf{k}(w_1 z_2 + w_2 z_1 + x_1 y_2 - x_2 y_1). \end{aligned} \tag{4.114}$$

Using the usual vector products in $\mathbb{R}^3$ (see 3.5.1.5, p. 184) it can be written in the form

$$q_1\,q_2 = (q_{01} + \underline{\mathbf{q}}_1)(q_{02} + \underline{\mathbf{q}}_2) = q_{01}\,q_{02} - \vec{\mathbf{q}}_1 \cdot \vec{\mathbf{q}}_2 + \vec{\mathbf{q}}_1 \times \vec{\mathbf{q}}_2\,, \tag{4.115}$$

where $\vec{\mathbf{q}}_1 \cdot \vec{\mathbf{q}}_2$ is the dot-product, and $\vec{\mathbf{q}}_1 \times \vec{\mathbf{q}}_2$ is the cross-product of the vectors $\vec{\mathbf{q}}_1, \vec{\mathbf{q}}_2 \in \mathbb{R}^3$. Next is to identify the $\mathbb{R}^3$ with the space $\mathbb{H}_0$ of the pure quaternions.

Remark: Multiplication of quaternions is not commutative!

The *product* q_1q_2 corresponds to the matrix multiplication of matrix $\mathbf{L}_{q_1}$ with vector q_2, and it is equal to the product of matrix $\mathbf{R}_{q_2}$ with q_1:

$$q_1q_2 = \mathbf{L}_{q_1}q_2 = \begin{pmatrix} w_1 & -x_1 & -y_1 & -z_1 \\ x_1 & w_1 & -z_1 & y_1 \\ y_1 & z_1 & w_1 & -x_1 \\ z_1 & -y_1 & x_1 & w_1 \end{pmatrix}\begin{pmatrix} w_2 \\ x_2 \\ y_2 \\ z_2 \end{pmatrix} = \mathbf{R}_{q_2}q_1 = \begin{pmatrix} w_2 & -x_2 & -y_2 & -z_2 \\ x_2 & w_2 & z_2 & -y_2 \\ y_2 & -z_2 & w_2 & x_2 \\ z_2 & y_2 & -x_2 & w_2 \end{pmatrix}\begin{pmatrix} w_1 \\ x_1 \\ y_1 \\ z_1 \end{pmatrix}. \tag{4.116}$$

3. Division

The definition of division of two quaternions is based on the multiplication: $q_1, q_2 \in \mathbb{H}$, $q_2 \neq 0$,

$$\frac{q_1}{q_2} := q_1\, q_2^{-1} = q_1 \frac{\bar{q}_2}{|q_2|^2}. \tag{4.117}$$

The order of the factors is important.

■ Let $q_1 = 1 + \mathbf{j}$, $q_2 = \frac{1}{\sqrt{2}}(1 - \mathbf{k})$, then $|q_2| = 1$, $\bar{q}_2 = \dfrac{1}{\sqrt{2}}(1 + \mathbf{k})$ and so

$$\frac{q_1}{q_2} := q_1 \frac{\bar{q}_2}{|q_2|^2} = \frac{1}{\sqrt{2}}(1 + \mathbf{i} + \mathbf{j} + \mathbf{k}) \neq \frac{\bar{q}_2}{|q_2|^2} q_1 = \frac{1}{\sqrt{2}}(1 - \mathbf{i} + \mathbf{j} + \mathbf{k}).$$

4. Generalized Moivre Formula

Let $q \in \mathbb{H}$, whrer $q = q_0 + \underline{\mathbf{q}} = r(\cos\varphi + \underline{\mathbf{n}}_{\underline{\mathbf{q}}} \sin\varphi)$ with $r = |q|$ and $\varphi = \arccos \dfrac{q_0}{|q|}$, $\cos\varphi = \dfrac{q_0}{|q|}$, $\sin\varphi = \dfrac{|\underline{\mathbf{q}}|}{|q|}$, then for arbitrary $k \in \mathbb{N}$:

$$q^k = r^k e^{\underline{\mathbf{n}}_{\underline{\mathbf{q}}} k\varphi} = r^k(\cos(k\varphi) + \underline{\mathbf{n}}_{\underline{\mathbf{q}}} \sin(k\varphi)). \tag{4.118}$$

5. Exponential Function

For $q = q_0 + \underline{\mathbf{q}} \in \mathbb{H}$ its exponential expression is defined as

$$e^q = \sum_{k=0}^{\infty} \frac{q^k}{k!} = e^{q_0}(\cos|\underline{\mathbf{q}}| + \underline{\mathbf{n}}_{\underline{\mathbf{q}}} \sin|\underline{\mathbf{q}}|). \tag{4.119}$$

Properties of the exponential function:

For any $q \in \mathbb{H}$ holds:

$$e^{-q}e^q = 1, \quad (4.120a) \qquad e^q \neq 0, \quad (4.120b) \qquad e^q = e^{q_0 + \underline{\mathbf{q}}} = e^{q_0}e^{\underline{\mathbf{q}}}, \quad (4.120c)$$

$$e^{\underline{\mathbf{n}}_{\underline{\mathbf{q}}}\pi} = -1, \text{ especially } e^{\mathbf{i}\pi} = e^{\mathbf{j}\pi} = e^{\mathbf{k}\pi} = -1. \tag{4.120d}$$

$$\text{Unit quaternion } u \text{ and } \vartheta \in \mathbb{R} : e^{\vartheta u} = \cos\vartheta + u\sin\vartheta. \tag{4.120e}$$

If $q_1\, q_2 = q_2\, q_1$ then $e^{q_1+q_2} = e^{q_1}\, e^{q_2}$. But it does not follow from $e^{q_1+q_2} = e^{q_1}\, e^{q_2}$ that $q_1\, q_2 = q_2\, q_1$.

■ Since $(\mathbf{i}\pi)(\mathbf{j}\pi) = \mathbf{k}\pi^2 \neq -\mathbf{k}\pi^2 = (\mathbf{j}\pi)(\mathbf{i}\pi)$ therefore also holds

$$e^{\mathbf{i}\pi}e^{\mathbf{j}\pi} = (\cos\pi)(\cos\pi) = (-1)(-1) = 1, \text{ but } e^{\mathbf{i}\pi+\mathbf{j}\pi} = \left(\cos\left(\sqrt{2}\pi\right) + \frac{\mathbf{i}+\mathbf{j}}{\sqrt{2}}\sin\left(\sqrt{2}\pi\right)\right) \neq 1.$$

6. Trigonometric Functions

For $q \in \mathbb{H}$ let

$$\cos q := \frac{1}{2}\left(e^{\underline{\mathbf{n}}_{\underline{\mathbf{q}}}q} + e^{-\underline{\mathbf{n}}_{\underline{\mathbf{q}}}q}\right), \qquad \sin q := -\underline{\mathbf{n}}_{\underline{\mathbf{q}}}\left(e^{\underline{\mathbf{n}}_{\underline{\mathbf{q}}}q} - e^{-\underline{\mathbf{n}}_{\underline{\mathbf{q}}}q}\right). \tag{4.121}$$

$\cos q$ is an even function, against which $\sin q$ is an odd function.

Addition formula: It is valid for any $q = q_0 + \underline{\mathbf{q}} \in \mathbb{H}$

$$\cos q = \cos q_0 \cos \underline{\mathbf{q}} - \sin q_0 \sin \underline{\mathbf{q}}\,, \quad \sin q = \sin q_0 \cos \underline{\mathbf{q}} + \cos q_0 \sin \underline{\mathbf{q}}\,. \tag{4.122}$$

7. Hyperbolic Functions

For $q \in \mathbb{H}$ let

$$\cosh q := \frac{1}{2}\left(e^q + e^{-q}\right), \qquad \sinh q := -\underline{\mathbf{n}}_{\underline{\mathbf{q}}}\left(e^q - e^{-q}\right). \tag{4.123}$$

$\cosh q$ is an even function, against which $\sinh q$ is an odd function.

Addition formula: It is valid for any $q = q_0 + \underline{\mathbf{q}} \in \mathbb{H}$

$$\cosh q = \cosh q_0 \cos \underline{\mathbf{q}} - \sinh q_0 \sinh \underline{\mathbf{q}}\,, \quad \sinh q = \sinh q_0 \cos \underline{\mathbf{q}} + \cosh q_0 \sinh \underline{\mathbf{q}}\,. \tag{4.124}$$

8. Logarithmic Function

For $q = q_0 + \underline{\mathbf{q}} = r(\cos\varphi + \underline{\mathbf{n}}_{\underline{\mathbf{q}}} \sin\varphi) \in \mathbb{H}$ and $k \in \mathbb{Z}$ the k-th branch of the logarithmic function is defined as

$$\log_k q := \begin{cases} \ln r + \underline{\mathbf{n}}_{\underline{\mathbf{q}}}(\varphi + 2k\pi), & |\underline{\mathbf{q}}| \neq 0 \quad \text{or} \quad |\underline{\mathbf{q}}| = 0 \text{ and } q_0 > 0\,, \\ \text{not defined for} & |\underline{\mathbf{q}}| = 0 \quad \text{and} \quad q_0 < 0. \end{cases} \tag{4.125}$$

Properties of the logarithmic function:

$$e^{\log_k q} = q \text{ for any } q \in \mathbb{H}\,, \text{ for which } \log_k q \text{ is defined}\,, \tag{4.126a}$$

$$\log_0 e^q = q \text{ for any } q \in \mathbb{H} \text{ with } |\underline{\mathbf{q}}| \neq (2l+1)\pi\,,\ l \in \mathbb{Z}\,, \tag{4.126b}$$

$$\log_k 1 = 0\,, \tag{4.126c}$$

$$\log_0 \mathbf{i} = \frac{\pi}{2}\mathbf{i}\,, \quad \log_0 \mathbf{j} = \frac{\pi}{2}\mathbf{j}, \quad \log_0 \mathbf{k} = \frac{\pi}{2}\mathbf{k}\,. \tag{4.126d}$$

In the case when $\log q_1$ and $\log q_2$ or q_1 and q_2 commute, then, if k is suitable defined, the following known equality (4.127) holds:

$$\log(q_1\, q_2) = \log q_1 + \log q_2\,. \tag{4.127}$$

For the unit quaternions $q \in \mathbb{H}_1$ holds $|q| = 1$ and $q = \cos\varphi + \underline{\mathbf{n}}_{\underline{\mathbf{q}}} \sin\varphi$ and so

$$\log q := \log_0 q = \underline{\mathbf{n}}_{\underline{\mathbf{q}}}\, \varphi \text{ for } q \neq -1\,, \tag{4.128}$$

9. Power function

Let $q \in \mathbb{H}$ and $\alpha \in \mathbb{R}$, then

$$q^\alpha := e^{\alpha \log q}. \tag{4.129}$$

4.4.2 Representation of Rotations in $\mathbb{R}^3$

Spatial rotations are performed around an axis, the so-called rotation axis. It goes through the origin. It is oriented by a direction vector $\vec{\mathbf{a}} \neq \vec{\mathbf{0}}$ (on the axis). The positive direction on this axis is chosen by $\vec{\mathbf{a}}$. The positive rotation (rotation angle $\varphi \geq 0$) is a counterclockwise rotation with respect to the positive direction. The direction vector is usually given normed, i.e. $|\vec{\mathbf{a}}| = 1$.

The equality

$$\vec{\mathbf{w}} = \mathbf{R}\vec{\mathbf{v}}\,, \tag{4.130a}$$

means vector $\vec{\mathbf{w}}$ arises from vector $\vec{\mathbf{v}}$ by the rotation matrix $\mathbf{R}$, i.e., the *rotation matrix* $\mathbf{R}$ transforms vector $\vec{\mathbf{v}}$ into $\vec{\mathbf{w}}$. Since rotation matrices are orthogonal matrices holds

$$\mathbf{R}^{-1} = \mathbf{R}^{\mathrm{T}} \tag{4.130b}$$

and (4.130a) is equivalent to

$$\vec{\mathbf{v}} = \mathbf{R}^{-1}\vec{\mathbf{w}} = \mathbf{R}^{\mathrm{T}}\vec{\mathbf{w}}. \tag{4.130c}$$

Remark: At spatial transformations it is necessary to distinguish between the followings:
a) *geometric transformations*, i.e. when geometric objects are transformed with respect to a fixed coordinate system, and
b) *coordinate transformations*, i.e. the object is fixed while the coordinate system is transformed with respect to the object (see 3.5.4, p. 229).
Now, geometric transformations are handled with quaternions.

4.4.2.1 Rotations of an Object About the Coordinate Axes

In a Cartesian coordinate system the axes are oriented by the basis vectors. The rotation around the x axis is given by matrix $\mathbf{R}_x$, around the y axis by $\mathbf{R}_y$ and around the z axis by $\mathbf{R}_z$, where:

$$\mathbf{R}_x(\alpha) := \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha & -\sin\alpha \\ 0 & \sin\alpha & \cos\alpha \end{pmatrix}, \quad \mathbf{R}_y(\beta) := \begin{pmatrix} \cos\beta & 0 & \sin\beta \\ 0 & 1 & 0 \\ -\sin\beta & 0 & \cos\beta \end{pmatrix}, \quad \mathbf{R}_z(\gamma) := \begin{pmatrix} \cos\gamma & -\sin\gamma & 0 \\ \sin\gamma & \cos\gamma & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{4.131}$$

The relation between a rotation of an object and the rotation of the coordinate system (see 3.5.3.3,**3.**, p. 213)) is

$$\mathbf{R}_x(\alpha) = \mathbf{D}_x^{\mathrm{T}}(\alpha)\,, \mathbf{R}_y(\beta) = \mathbf{D}_y^{\mathrm{T}}(\beta)\,, \mathbf{R}_z(\gamma) = \mathbf{D}_z^{\mathrm{T}}(\gamma)\,. \tag{4.132}$$

Remark: The representation of the rotation matrices in homogenous coordinates is given in 3.5.4.5,**1.**,**2.**, p. 234.

4.4.2.2 Cardan-Angles

Every rotation $\mathbf{R}$ around an axis passing through the origin can be given as a sequence of rotations around the coordinate axes in a given coordinate system (see also 3.5.3.5, p. 214), where here
• the first rotation is around the x axis by an angle α_C, then
• the second rotation is around the y axis by an angle β_C, then
• the third rotation is around the z axis by an angle γ_C.
The angles α_C, β_C, γ_C are called Cardan-angles. Then the rotation matrix is

$$\mathbf{R} = \mathbf{R}_{\mathrm{C}} := \mathbf{R}_z(\gamma_{\mathrm{C}})\mathbf{R}_y(\beta_{\mathrm{C}})\mathbf{R}_x(\alpha_{\mathrm{C}}) \tag{4.133a}$$

$$= \begin{pmatrix} \cos\beta_{\mathrm{C}}\cos\gamma_{\mathrm{C}} & \sin\alpha_{\mathrm{C}}\sin\beta_{\mathrm{C}}\cos\gamma_{\mathrm{C}} - \cos\alpha_{\mathrm{C}}\sin\gamma_{\mathrm{C}} & \cos\alpha_{\mathrm{C}}\sin\beta_{\mathrm{C}}\cos\gamma_{\mathrm{C}} + \sin\alpha_{\mathrm{C}}\sin\gamma_{\mathrm{C}} \\ \cos\beta_{\mathrm{C}}\sin\gamma_{\mathrm{C}} & \sin\alpha_{\mathrm{C}}\sin\beta_{\mathrm{C}}\sin\gamma_{\mathrm{C}} + \cos\alpha_{\mathrm{C}}\cos\gamma_{\mathrm{C}} & \cos\alpha_{\mathrm{C}}\sin\beta_{\mathrm{C}}\sin\gamma_{\mathrm{C}} - \sin\alpha_{\mathrm{C}}\cos\gamma_{\mathrm{C}} \\ -\sin\beta_{\mathrm{C}} & \sin\alpha_{\mathrm{C}}\cos\beta_{\mathrm{C}} & \cos\alpha_{\mathrm{C}}\cos\beta_{\mathrm{C}} \end{pmatrix}. \tag{4.133b}$$

Adventages:
• very popular representation of rotations,
• clear structure.

Disadventages:
• the order of rotations is important i.e. in general holds

$$\mathbf{R}_x(\alpha_{\mathrm{C}})\mathbf{R}_y(\beta_{\mathrm{C}})\mathbf{R}_z(\gamma_{\mathrm{C}}) \neq \mathbf{R}_z(\gamma_{\mathrm{C}})\mathbf{R}_y(\beta_{\mathrm{C}})\mathbf{R}_x(\alpha_{\mathrm{C}})\,, \tag{4.133c}$$

• the representation is not unique since $\mathbf{R}(\alpha_{\mathrm{C}},\, \beta_{\mathrm{C}},\, \gamma_{\mathrm{C}}) = \mathbf{R}(-\alpha_{\mathrm{C}} \pm 180°,\, \beta_{\mathrm{C}} \pm 180°,\, \gamma_{\mathrm{C}} \pm 180°)\,,$
• not suitable for rotations after each other (e.g. at animations),
• gimbal lock can happen (rotation of an axis by 90° goes into an other axis)

■ Gimbal Lock case: rotation around the y axis by 90°

$$\mathbf{R}(\alpha_{\mathrm{C}}, 90°, \gamma_{\mathrm{C}}) = \begin{pmatrix} 0 & \sin(\alpha_{\mathrm{C}} - \gamma_{\mathrm{C}}) & \cos(\alpha_{\mathrm{C}} - \gamma_{\mathrm{C}}) \\ 0 & \cos(\alpha_{\mathrm{C}} - \gamma_{\mathrm{C}}) & -\sin(\alpha_{\mathrm{C}} - \gamma_{\mathrm{C}}) \\ -1 & 0 & 0 \end{pmatrix}. \tag{4.133d}$$

It can be seen that one degree of freedom is lost. In practical applications it can lead to unpredictable motions.

Remark: It can be realized that Cardan-angles are called sometimes as Euler-angles, in the literature however their definitions can be different (see 3.5.3.6, p. 215).

4.4.2.3 Euler Angles

The Euler-angles ψ, ϑ, φ are often introduced as follows (see 3.5.3.6, p. 215):
• the first rotation around the z axis by angle ψ,
• the second rotation around the image of x axis by angle ϑ,
• the third rotation around the image of z axis by angle φ.
The rotation matrix is

$$\mathbf{R} = \mathbf{R}_{\mathrm{E}} := \mathbf{R}_z(\varphi)\mathbf{R}_x(\vartheta)\mathbf{R}_z(\psi) \tag{4.134a}$$

$$= \begin{pmatrix} \cos\psi\cos\varphi - \sin\psi\cos\vartheta\sin\varphi & -\cos\psi\sin\varphi - \sin\psi\cos\vartheta\cos\varphi & \sin\psi\sin\vartheta \\ \sin\psi\cos\varphi + \cos\psi\cos\vartheta\sin\varphi & -\sin\psi\sin\varphi + \cos\psi\cos\vartheta\cos\varphi & -\cos\psi\sin\vartheta \\ \sin\vartheta\sin\varphi & \sin\vartheta\cos\varphi & \cos\vartheta \end{pmatrix}. \tag{4.134b}$$

4.4.2.4 Rotation Around an Arbitrary Zero Point Axis

The counterclockwise rotation around a normed vector $\vec{\mathbf{a}} = (a_x,\, a_y,\, a_z)$ with $|\vec{\mathbf{a}}| = 1$ by an angle φ is made in 5 steps:

1. Rotation of $\vec{\mathbf{a}}$ around the y axis until reaching the x, y plane: $\vec{\mathbf{a}}' = \mathbf{R}_1\vec{\mathbf{a}}$ using $\mathbf{R}_1$ according to (4.135a). The result: vector $\vec{\mathbf{a}}'$ is in the x, y plane.

2. Rotation of $\vec{\mathbf{a}}'$ around the z axis until becoming parallel to the x axis $\vec{\mathbf{a}}'' = \mathbf{R}_2\vec{\mathbf{a}}'$ by $\mathbf{R}_2$ according to (4.135b). The result: vector $\vec{\mathbf{a}}''$ is parallel to the x axis.

$$\mathbf{R}_1 = \begin{pmatrix} \frac{a_x}{\sqrt{a_x^2+a_z^2}} & 0 & \frac{a_z}{\sqrt{a_x^2+a_z^2}} \\ 0 & 1 & 0 \\ \frac{-a_z}{\sqrt{a_x^2+a_z^2}} & 0 & \frac{a_x}{\sqrt{a_x^2+a_z^2}} \end{pmatrix}, \tag{4.135a} \qquad \mathbf{R}_2 = \begin{pmatrix} \sqrt{a_x^2+a_z^2} & a_y & 0 \\ -a_y & \sqrt{a_x^2+a_z^2} & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{4.135b}$$

3. Rotation around the x axis by an angle φ by $\mathbf{R}_3$:

$$\mathbf{R}_3 = \mathbf{R}_x(\varphi) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\varphi & -\sin\varphi \\ 0 & \sin\varphi & \cos\varphi \end{pmatrix}. \tag{4.135c}$$

Rotations $\mathbf{R}_1$ and $\mathbf{R}_2$ are inverted in the following two steps.

4. Inverse rotation of $\mathbf{R}_2$, i.e. rotation around the z axis by the angle β where $\sin\beta = a_y$, $\cos\beta = \sqrt{a_x^2+a_z^2}$ according to (4.135d).

5. Inverse rotation of $\mathbf{R}_1$, i.e. rotation around the y axis by $-\alpha$ where $\sin(-\alpha) = \frac{-a_z}{\sqrt{a_x^2+a_z^2}}$, $\cos(-\alpha) = \frac{a_x}{\sqrt{a_x^2+a_z^2}}$ around the y axis according to (4.135e).

$$\mathbf{R}_2^{-1} = \begin{pmatrix} \sqrt{a_x^2+a_z^2} & -a_y & 0 \\ a_y & \sqrt{a_x^2+a_z^2} & 0 \\ 0 & 0 & 1 \end{pmatrix}, \tag{4.135d} \qquad \mathbf{R}_1^{-1} = \begin{pmatrix} \frac{a_x}{\sqrt{a_x^2+a_z^2}} & 0 & \frac{-a_z}{\sqrt{a_x^2+a_z^2}} \\ 0 & 1 & 0 \\ \frac{a_z}{\sqrt{a_x^2+a_z^2}} & 0 & \frac{a_x}{\sqrt{a_x^2+a_z^2}} \end{pmatrix}. \tag{4.135e}$$

Finally the composition matrix is:

$$\mathbf{R}(\vec{\mathbf{a}},\, \varphi) = \mathbf{R}_1^{-1}\mathbf{R}_2^{-1}\mathbf{R}_3\mathbf{R}_2\mathbf{R}_1 = \tag{4.135f}$$

$$\begin{pmatrix} \cos\varphi + a_x^2(1-\cos\varphi) & a_x a_y(1-\cos\varphi) - a_z\sin\varphi & a_x a_z(1-\cos\varphi) + a_y\sin\varphi \\ a_y a_x(1-\cos\varphi) + a_z\sin\varphi & \cos\varphi + a_y^2(1-\cos\varphi) & a_y a_z(1-\cos\varphi) - a_x\sin\varphi \\ a_z a_x(1-\cos\varphi) - a_y\sin\varphi & a_z a_y(1-\cos\varphi) + a_x\sin\varphi & \cos\varphi + a_z^2(1-\cos\varphi) \end{pmatrix}. \quad (4.135g)$$

Matrix $\mathbf{R}(\vec{\mathbf{a}}, \varphi)$ is an orthogonal matrix, i.e. its inverse is equal to its transpose: $\mathbf{R}^{-1}(\vec{\mathbf{a}}, \varphi) = \mathbf{R}^{\mathrm{T}}(\vec{\mathbf{a}}, \varphi)$. The following formulas are also valid:

$$\mathbf{R}\vec{\mathbf{x}} = \mathbf{R}(\vec{\mathbf{a}}, \varphi)\vec{\mathbf{x}}$$

$$= (\cos\varphi)\vec{\mathbf{x}} + (1-\cos\varphi)\frac{\vec{\mathbf{x}}\cdot\vec{a}}{|\vec{\mathbf{a}}|^2}\vec{\mathbf{a}} + \frac{\sin\varphi}{|a|}\vec{\mathbf{a}}\times\vec{\mathbf{x}} \quad (4.136a)$$

$$= (\cos\varphi)\vec{\mathbf{x}} + (1-\cos\varphi)\vec{\mathbf{x}}_{\vec{\mathbf{a}}} + (\sin\varphi)\frac{\vec{\mathbf{a}}}{|\vec{\mathbf{a}}|}\times\vec{\mathbf{x}}. \quad (4.136b)$$

In these formulas the vector $\vec{\mathbf{x}}$ is decomposed into two components, one is parallel, the other is perpendicular to $\vec{\mathbf{a}}$. The parallel part is $\vec{\mathbf{x}}_{\vec{\mathbf{a}}} = \frac{\vec{\mathbf{x}}\cdot\vec{\mathbf{a}}}{|\vec{\mathbf{a}}|^2}\vec{\mathbf{a}}$, the perpendicular part is $\vec{\mathbf{r}} = \vec{\mathbf{x}} - \vec{\mathbf{x}}_{\vec{\mathbf{a}}}$. The orthogonal part is in a plane whose normal vector is $\vec{\mathbf{a}}$, so its image is $\cos\varphi\vec{\mathbf{r}} + \sin\varphi\vec{\mathbf{r}}^*$, where $\vec{\mathbf{r}}^*$ is obtained from $\vec{\mathbf{r}}$ by a 90° rotation in positive direction: $\vec{\mathbf{r}}^* = \frac{1}{|\vec{\mathbf{a}}|}\vec{\mathbf{a}}\times\vec{\mathbf{r}}$. The result of the rotation of vector $\vec{\mathbf{x}}$ is

$$\vec{\mathbf{x}}_{\vec{\mathbf{a}}} + \cos\varphi\vec{\mathbf{r}} + \sin\varphi\vec{\mathbf{r}}^* = \frac{\vec{\mathbf{x}}\cdot\vec{\mathbf{a}}}{|\vec{\mathbf{a}}|^2}\vec{\mathbf{a}} + (\cos\varphi)\left(\vec{\mathbf{x}} - \frac{\vec{\mathbf{x}}\cdot\vec{\mathbf{a}}}{|\vec{\mathbf{a}}|^2}\vec{\mathbf{a}}\right) + (\sin\varphi)\frac{1}{|\vec{\mathbf{a}}|}\vec{\mathbf{a}}\times\vec{\mathbf{r}} \quad (4.136c)$$

$$\text{with } \vec{\mathbf{a}}\times\vec{\mathbf{r}} = \vec{\mathbf{a}}\times(\vec{\mathbf{x}} - \vec{\mathbf{x}}_a) = \vec{\mathbf{a}}\times\vec{\mathbf{x}}. \quad (4.136d)$$

Advantages:
- „Standard representation" in computer-graphics,
- CARDAN-angles should not be determined,
- no gimbal lock.

Disadvantage: Not suitable for animation, i.e. for interpolation of rotations.

4.4.2.5 Rotation and Quaternions

If the unit vector $\vec{\mathbf{a}}$ in (4.135f) is identified as the pure quaternion $\underline{\mathbf{a}}$ (while the angle of rotation φ remains the same), then one gets:

$$R(\underline{\mathbf{a}}, \varphi) = \begin{pmatrix} q_0^2+q_1^2-q_2^2-q_3^2 & 2q_1q_2-2q_0q_3 & 2q_1q_3+2q_0q_2 \\ 2q_1q_2+2q_0q_3 & q_0^2-q_1^2+q_2^2-q_3^2 & 2q_2q_3-2q_0q_1 \\ 2q_1q_3-2q_0q_2 & 2q_2q_3+2q_0q_1 & q_0^2-q_1^2-q_2^2+q_3^2 \end{pmatrix} =: \mathbf{R}(q) \quad (4.137a)$$

where $q_0 = \cos\frac{\varphi}{2}$ and $\underline{\mathbf{q}} = (q_1, q_2, q_3)^{\mathrm{T}} = (a_x, a_y, a_z)^{\mathrm{T}}\sin\frac{\varphi}{2}$, i.e. q is the unit quaternion $q = q(\underline{\mathbf{a}}, \varphi) = \cos\frac{\varphi}{2} + \underline{\mathbf{a}}\sin\frac{\varphi}{2} \in \mathbb{H}_1$. If vector $\vec{\mathbf{x}}$ is considered as $\mathbb{R}^3 \ni \vec{\mathbf{x}} = x_1\mathbf{i} + x_2\mathbf{j} + x_3\mathbf{k} \in \mathbb{H}_0$, then

$$\mathbf{R}(\underline{\mathbf{a}}, \varphi)\,\underline{\mathbf{x}} = \mathbf{R}(q)\,\underline{\mathbf{x}} = q\,\underline{\mathbf{x}}\,\bar{q}. \quad (4.137b)$$

Especially the columns of the rotation matrix are vectors $q\,\underline{\mathbf{e}}_k\,\bar{q}$:

$$\mathbf{R}(\underline{\mathbf{a}}, \varphi) = \left(q\begin{pmatrix}1\\0\\0\end{pmatrix}\bar{q} \quad q\begin{pmatrix}0\\1\\0\end{pmatrix}\bar{q} \quad q\begin{pmatrix}0\\0\\1\end{pmatrix}\bar{q} \right) = (\, q\,\mathbf{i}\,\bar{q} \quad q\,\mathbf{j}\,\bar{q} \quad q\,\mathbf{k}\,\bar{q}\,). \quad (4.137c)$$

Consequences:

- The matrix of rotation can be determined with the help of quaternion $q = \cos\frac{\varphi}{2} + \underline{\mathbf{a}}\sin\frac{\varphi}{2}$.

• For the rotated vector $\mathbf{R}(\underline{\mathbf{a}},\,\varphi)\underline{\mathbf{x}}$ holds $\mathbf{R}(\underline{\mathbf{a}},\,\varphi)\underline{\mathbf{x}} = q\,\underline{\mathbf{x}}\,\overline{q}$ in the sense of quaternion multiplication and identifying $\mathbb{R}^3$ with the set of pure quaternions $\mathbb{H}_0$.

For every unit quaternion $q \in \mathbb{H}_1$ q and $-q$ determine the same rotation, so $\mathbb{H}_1$ is a double covering of $SO(3)$. Performing rotations one after the other corresponds to multiplication of quaternions, i.e.

$$\mathbf{R}(q_2)\,\mathbf{R}(q_1) = \mathbf{R}(q_2\,q_1)\,; \tag{4.138}$$

and the *conjugate quaternion* corresponds to the inverse rotation:

$$\mathbf{R}^{-1}(q) = \mathbf{R}(\overline{q}). \tag{4.139}$$

■ Rotation by 60° around the axis defined by $(1,1,1)^{\mathrm{T}}$. First the direction vector should be normed: $\underline{\mathbf{a}} = \frac{1}{\sqrt{3}}(1,1,1)^{\mathrm{T}}$. Then with $\sin\varphi = \sin 60^\circ = \frac{\sqrt{3}}{2}$ and $\cos\varphi = \cos 60^\circ = \frac{1}{2}$ the rotation matrix becomes

$$\mathbf{R}\left(\frac{1}{\sqrt{3}}(1,1,1)^{\mathrm{T}},\,60^\circ\right) = \frac{1}{3}\begin{pmatrix} 2 & -1 & 2 \\ 2 & 2 & -1 \\ -1 & 2 & 2 \end{pmatrix}.$$

The quaternion describing the rotation is:

$$\begin{aligned} q = q\left(\frac{1}{\sqrt{3}}(1,1,1)^{\mathrm{T}},\,60^\circ\right) &= \cos 30^\circ + \frac{1}{\sqrt{3}}(\mathbf{i}+\mathbf{j}+\mathbf{k})\sin 30^\circ \\ &= \frac{\sqrt{3}}{2} + \frac{1}{\sqrt{3}}(\mathbf{i}+\mathbf{j}+\mathbf{k})\frac{1}{2} = \frac{\sqrt{3}}{2} + \frac{\sqrt{3}}{6}(\mathbf{i}+\mathbf{j}+\mathbf{k})\,. \end{aligned}$$

Furthermore

$$\begin{aligned} q\begin{pmatrix}1\\0\\0\end{pmatrix}\overline{q} &= \left(\frac{\sqrt{3}}{2} + \frac{\sqrt{3}}{6}(\mathbf{i}+\mathbf{j}+\mathbf{k}\right)\mathbf{i}\left(\frac{\sqrt{3}}{2} - \frac{\sqrt{3}}{6}(\mathbf{i}+\mathbf{j}+\mathbf{k})\right) \\ &= \left(\frac{\sqrt{3}}{2} + \frac{\sqrt{3}}{6}(\mathbf{i}+\mathbf{j}+\mathbf{k})\right)\left(\frac{\sqrt{3}}{2}\mathbf{i} + \frac{\sqrt{3}}{6} - \frac{\sqrt{3}}{6}\mathbf{k} + \frac{\sqrt{3}}{6}\mathbf{j}\right) \\ &= \frac{24}{36}\mathbf{i} + \frac{24}{36}\mathbf{j} - \frac{12}{36}\mathbf{k} = \frac{1}{3}(2\mathbf{i}+2\mathbf{j}-\mathbf{k}) \triangleq \frac{1}{3}\begin{pmatrix}2\\2\\-1\end{pmatrix}. \end{aligned}$$

The two other columns are determined analogously:

$$q\begin{pmatrix}0\\1\\0\end{pmatrix}\overline{q} = \left(\frac{\sqrt{3}}{2} + \frac{\sqrt{3}}{6}(\mathbf{i}+\mathbf{j}+\mathbf{k}\right)\mathbf{j}\left(\frac{\sqrt{3}}{2} - \frac{\sqrt{3}}{6}(\mathbf{i}+\mathbf{j}+\mathbf{k})\right) = \frac{1}{3}(-\mathbf{i}+2\mathbf{j}+2\mathbf{k}) \triangleq \frac{1}{3}\begin{pmatrix}-1\\2\\2\end{pmatrix},$$

$$q\begin{pmatrix}0\\0\\1\end{pmatrix}\overline{q} = \left(\frac{\sqrt{3}}{2} + \frac{\sqrt{3}}{6}(\mathbf{i}+\mathbf{j}+\mathbf{k}\right)\mathbf{k}\left(\frac{\sqrt{3}}{2} - \frac{\sqrt{3}}{6}(\mathbf{i}+\mathbf{j}+\mathbf{k})\right) = \frac{1}{3}(2\mathbf{i}-\mathbf{j}+2\mathbf{k}) \triangleq \frac{1}{3}\begin{pmatrix}2\\-1\\2\end{pmatrix},$$

$$\mathbf{R}\left(\frac{1}{\sqrt{3}}\begin{pmatrix}1\\1\\1\end{pmatrix},\,60^\circ\right) = \left(q\begin{pmatrix}1\\0\\0\end{pmatrix}\overline{q}\quad q\begin{pmatrix}0\\1\\0\end{pmatrix}\overline{q}\quad q\begin{pmatrix}0\\0\\1\end{pmatrix}\overline{q}\right) = \frac{1}{3}\begin{pmatrix} 2 & -1 & 2 \\ 2 & 2 & -1 \\ -1 & 2 & 2 \end{pmatrix}.$$

4.4.2.6 Quaternions and Cardan Angles

The rotation matrix in Cardan angles (see (4.133a,b), p. 295) is exactly a matrix of rotation with a unit quaternion $q \in \mathbb{H}_1$.

$$\mathbf{R}_{\mathrm{C}}(\alpha_{\mathrm{C}},\,\beta_{\mathrm{C}},\,\gamma_{\mathrm{C}}) = \mathbf{R}_z(\gamma_{\mathrm{C}})\mathbf{R}_y(\beta_{\mathrm{C}})\mathbf{R}_x(\alpha_{\mathrm{C}}) \tag{4.140a}$$

$$= \begin{pmatrix} \cos\beta_C \cos\gamma_C & \sin\alpha_C \sin\beta_C \cos\gamma_C - \cos\alpha_C \sin\gamma_C & \cos\alpha_C \sin\beta_C \cos\gamma_C + \sin\alpha_C \sin\gamma_C \\ \cos\beta_C \sin\gamma_C & \sin\alpha_C \sin\beta_C \sin\gamma_C + \cos\alpha_C \cos\gamma_C & \cos\alpha_C \sin\beta_C \sin\gamma_C - \sin\alpha_C \cos\gamma_C \\ -\sin\beta_C & \sin\alpha_C \cos\beta_C & \cos\alpha_C \cos\beta_C \end{pmatrix} \quad (4.140b)$$

$$= [r_{ij}]_{i,j=1}^3 = \begin{pmatrix} q_0^2 + q_1^2 - q_2^2 - q_3^2 & 2q_1q_2 - 2q_0q_3 & 2q_1q_3 + 2q_0q_2 \\ 2q_1q_2 + 2q_0q_3 & q_0^2 - q_1^2 + q_2^2 - q_3^2 & 2q_2q_3 - 2q_0q_1 \\ 2q_1q_3 - 2q_0q_2 & 2q_2q_3 + 2q_0q_1 & q_0^2 - q_1^2 - q_2^2 + q_3^2 \end{pmatrix} = \mathbf{R}(q) \quad (4.140c)$$

$$= \left(q \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \overline{q} \quad q \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \overline{q} \quad q \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \overline{q} \right). \quad (4.140d)$$

Comparing the matrix elements one gets

$$\tan\gamma_C = \frac{r_{21}}{r_{11}}, \quad \sin\beta_C = -r_{31}, \quad \tan\alpha_C = \frac{r_{32}}{r_{33}}. \quad (4.141a)$$

In general, the solution is not unique, which is typical in trigonometric problems. However the uniqueness of angles can be reached by discussion of the defined domains.
Reversed, it is easy to get the unit quaternion from the rotation matrix.

$$4q_0q_1 = r_{32} - r_{23}, \quad 4q_0q_2 = r_{13} - r_{31}, \quad 4q_0q_3 = r_{21} - r_{12}\,, \quad (4.141b)$$

$$4q_0^2 - 1 = 4q_0^2 - q_0^2 - q_1^2 - q_2^2 - q_3^2 = r_{11} + r_{22} + r_{33}. \quad (4.141c)$$

Since q and $-q$ define the same rotation, q_0 can be determined as

$$q_0 = \frac{1}{2}\sqrt{r_{11} + r_{22} + r_{33} + 1}\,. \quad (4.141d)$$

The other components are

$$q_1 = \frac{r_{32} - r_{23}}{4q_0}, \quad q_2 = \frac{r_{13} - r_{31}}{4q_0}, \quad q_3 = \frac{r_{21} - r_{12}}{4q_0}. \quad (4.141e)$$

■ Let the rotation matrix be the following:

$$\mathbf{R} = \frac{1}{2} \begin{pmatrix} \sqrt{2} & -\frac{1}{2}\sqrt{6} & \frac{1}{2}\sqrt{2} \\ \sqrt{2} & \frac{1}{2}\sqrt{6} & -\frac{1}{2}\sqrt{2} \\ 0 & 1 & \sqrt{3} \end{pmatrix}.$$

1. Determination of the Cardan angles: Based on the above formulas $\sin\beta_C = -r_{31} = 0$, so $\beta_C = k\pi$, $k \in \mathbb{Z}$. Furthermore $\tan\gamma_C = \frac{r_{21}}{r_{11}} = 1$, so $\gamma_C = \frac{\pi}{4} + k\pi$, $k \in \mathbb{Z}$, and from $\tan\alpha_C = \frac{r_{32}}{r_{33}} = \frac{1}{\sqrt{3}}$ it follows that $\alpha_C = \frac{\pi}{6} + k\pi$, $k \in \mathbb{Z}$. The angles are unique, if they are determined as the ,,possible smallest" ones, i.e. the rotation whose angles have an absolute value $\le \frac{\pi}{2}$. So the angles are

$$\alpha_C = \frac{\pi}{6}, \quad \beta_C = 0, \quad \gamma_C = \frac{\pi}{4}.$$

2. Determination of the unit quaternion which results this rotation:

$$4q_0^2 - 1 = \frac{1}{2}\left(\sqrt{2} + \frac{1}{2}\sqrt{6} + \sqrt{3}\right) \quad \text{also} \quad q_0 = \frac{1}{2}\sqrt{1 + \frac{1}{2}(\sqrt{2} + \frac{1}{2}\sqrt{6} + \sqrt{3})} \approx 0,8924 = \cos\frac{\varphi}{2}\,.$$

The (possible smallest) angle is $\varphi = 53.6474°$, so $\sin\frac{\varphi}{2} = 0.4512$.

3. Determination of the further components of q and the direction of the axis of rotation

$\underline{\mathbf{a}} = (a_x,\, a_y,\, a_z)^{\mathrm{T}}$:

$$q_1 = \frac{r_{32}-r_{23}}{4q_0} = \frac{\left(\frac{1}{2}+\frac{1}{4}\sqrt{2}\right)}{4q_0} \approx 0,2391 \qquad \text{so } a_x = \frac{q_1}{\sin\frac{\varphi}{2}} \approx 0,5299,$$

$$q_2 = \frac{r_{13}-r_{31}}{4q_0} = \frac{\frac{1}{2}\cdot\frac{1}{2}\sqrt{2}}{4q_0} \approx 0,0991 \qquad \text{so } a_y = \frac{q_2}{\sin\frac{\varphi}{2}} \approx 0,2195,$$

$$q_3 = \frac{r_{21}-r_{12}}{4q_0} = \frac{\frac{1}{2}\left(\sqrt{2}+\frac{1}{2}\sqrt{6}\right)}{4q_0} \approx 0,3696 \text{ so } a_z = \frac{q_3}{\sin\frac{\varphi}{2}} \approx 0,8192.$$

Remark: At the calculation of the components in (4.141e) it can be a problem when q_0 is zero or close to zero. In this case the unit quaternion can not be determined by the formulas in (4.141e). To understand this situation one discusses the trace of the rotation matrix:

$$\operatorname{Tr}\mathbf{R} = r_{11}+r_{22}+r_{33} = 4q_0^2-1\,. \tag{4.142a}$$

If $\operatorname{Tr}\mathbf{R} > 0$, then $q_0 = \frac{1}{2}\sqrt{\operatorname{Tr}\mathbf{R}+1} > 0$, and the formulas (4.141e) can be used without any problem. If $\operatorname{Tr}\mathbf{R} \le 0$, then q_0 can be close to zero. In this case the greatest element of the main diagonal is considered. Assume, it is r_{11}. Then $|q_1|$ is greater than $|q_2|$ or $|q_3|$. The components q_1, q_2, q_3 also can be determined from the elements of the main diagonal of the rotation matrix. Choosing the positive sign for the square-roots follows:

$$q_1 = \frac{1}{2}\sqrt{1+r_{11}-r_{22}-r_{33}}\,,\quad q_2 = \frac{1}{2}\sqrt{1+r_{22}-r_{11}-r_{33}}\,,\quad q_3 = \frac{1}{2}\sqrt{1+r_{33}-r_{11}-r_{22}}\,. \tag{4.142b}$$

Calculation rules: From this facts the following calculation rules are derived:

- If $\operatorname{Tr}\mathbf{R} \le 0$ and $r_{11} \ge r_{22}$ and $r_{11} \ge r_{33}$, then q_1 has the greatest absolute value, so

$$q_0 = \frac{r_{32}-r_{23}}{4q_1},\quad q_2 = \frac{r_{21}+r_{12}}{4q_1},\quad q_3 = \frac{r_{13}+r_{31}}{4q_1}\,. \tag{4.142c}$$

- If $\operatorname{Tr}\mathbf{R} \le 0$ and $r_{22} \ge r_{11}$ and $r_{22} \ge r_{33}$, then q_2 has the greatest absolute value, so

$$q_0 = \frac{r_{13}-r_{31}}{4q_2},\quad q_1 = \frac{r_{21}+r_{12}}{4q_2},\quad q_3 = \frac{r_{23}+r_{32}}{4q_2}\,. \tag{4.142d}$$

- If $\operatorname{Tr}\mathbf{R} \le 0$ and $r_{33} \ge r_{11}$ and $r_{33} \ge r_{22}$, then q_3 has the greatest absolute value, so

$$q_0 = \frac{r_{21}-r_{12}}{4q_3},\quad q_1 = \frac{r_{31}+r_{13}}{4q_3},\quad q_2 = \frac{r_{23}+r_{32}}{4q_3}\,. \tag{4.142e}$$

Since the CARDAN-angles define the rotations around the corresponding axes, one can find the assignments given in the following table. Then the rotation

$$\mathbf{R}(\alpha,\,\beta,\,\gamma) = \mathbf{R}((0,0,1)^{\mathrm{T}},\,\gamma)\mathbf{R}((0,1,0)^{\mathrm{T}},\,\beta)\mathbf{R}((1,0,0)^{\mathrm{T}},\,\alpha) \tag{4.142f}$$

corresponds to the unit quaternion

$$q = Q_z\,Q_y\,Q_x\,. \tag{4.142g}$$

rotation	Cardan angle	around	quaternion
$\mathbf{R}_C((1,0,0)^T,\ \alpha_C)$	α_C	x axis	$Q_x := \cos\frac{\alpha_C}{2} + \mathbf{i}\sin\frac{\alpha_C}{2}$
$\mathbf{R}_C((0,1,0)^T,\ \beta_C)$	β_C	y axis	$Q_y := \cos\frac{\beta_C}{2} + \mathbf{j}\sin\frac{\beta_C}{2}$
$\mathbf{R}_C((0,0,1)^T,\ \gamma_C)$	γ_C	z axis	$Q_z := \cos\frac{\gamma_C}{2} + \mathbf{k}\sin\frac{\gamma_C}{2}$

■ Knowing the Cardan angles $\alpha_C = \frac{\pi}{6}$, $\beta_C = 0$, $\gamma_C = \frac{\pi}{4}$, the quaternion describing this rotation can be determined in the following way:

$$Q_x = \cos\frac{\alpha_C}{2} + \mathbf{i}\sin\frac{\alpha_C}{2} = \cos\frac{\pi}{12} + \mathbf{i}\sin\frac{\pi}{12},$$

$$Q_y = \cos\frac{\beta_C}{2} + \mathbf{j}\sin\frac{\beta_C}{2} = \cos 0 + \mathbf{j}\sin 0 = 1,$$

$$Q_z = \cos\frac{\gamma_C}{2} + \mathbf{k}\sin\frac{\gamma_C}{2} = \cos\frac{\pi}{8} + \mathbf{k}\sin\frac{\pi}{8}.$$

The final result coincides with that given on page 299:

$$\begin{aligned} q := Q_z\,Q_y\,Q_x &= \left(\cos\frac{\pi}{8} + \mathbf{k}\sin\frac{\pi}{8}\right) 1 \left(\cos\frac{\pi}{12} + \mathbf{i}\sin\frac{\pi}{12}\right) \\ &= \cos\frac{\pi}{8}\cdot\cos\frac{\pi}{12} + \mathbf{i}\cos\frac{\pi}{8}\cdot\sin\frac{\pi}{12} + \mathbf{j}\sin\frac{\pi}{8}\cdot\sin\frac{\pi}{12} + \mathbf{k}\sin\frac{\pi}{8}\cdot\cos\frac{\pi}{12} \\ &= 0,8924 + 0,2391\mathbf{i} + 0,0991\mathbf{j} + 0,3696\mathbf{k}. \end{aligned}$$

4.4.2.7 Efficiency of the Algorithms

To estimate the efficiency of the algorithms standard operations are defined from which the more complex operations are originated. For complicated comparisons with other methods see [4.12].

Let

- M: number of **mu**ltiplications,
- A: number of **a**dditions and subtractions,
- D: number of **d**ivisions,
- S: number of **s**tandard functions calls, e.g. trigonometric functions, which are composed of a considerable number of multiplications, divisions and additions,
- C: number of **c**omparisons of expressions, which increase the computing time by interrupting the algorithm.

Operation	A	M	D	S	C
Quaternion into Matrix	12	12			
Matrix into Quaternion ($\mathrm{Tr}\,\mathbf{R} > \mathbf{0}$)	6	5	1	1	1
Matrix into Quaternion ($\mathrm{Tr}\,\mathbf{R} \le \mathbf{0}$)	6	5	1	1	3

Rotation of a vector	A	M	Remarks
with rotation matrix	6	9	
with unit quaternion	24	32	normal quaternion multiplication
with unit quaternion	17	24	fast quaternion multiplication
with unit quaternion	18	21	changing into rotation matrix

Rotation of n vectors	A	M	Remarks
with rotation matrix	$6n$	$9n$	
with unit quaternion	$24n$	$32n$	normal quaternion multiplication
with unit quaternion	$17n$	$24n$	fast quaternion multiplication
with unit quaternion	$12+6n$	$12+9n$	changing into rotation matrix

composition of two rotations	A	M
with rotation matrix	18	27
with unit quaternion	12	16

Summary: An algorithm based on quaternions is faster only when rotations are performed after each other. It occurs mainly in computer graphics at animations, i.e. at approximations of rotations.

4.4.3 Applications of Quaternions

4.4.3.1 3D Rotations in Computer Graphics

To describe motion flows interpolation of rotations are used. Since the 3D-rotations can be represented by unit quaternions, algorithms are developed for interpolation of rotations in computer graphics. The easiest idea is to start analogously to the definition of linear interpolation in Euclidian-spaces. Basic algorithms are Lerp, Slerp and Squad.

1. Lerp (linear interpolation)

Let $p, q \in \mathbb{H}_1$ and $t \in [0, 1]$, then

$$\mathrm{Lerp}(p,\, q,\, t) = p(1-t) + q\,t\,. \tag{4.143}$$

• This is a linear segment in $\mathbb{R}^4$, connecting $p \in \mathbb{H}_1 \sim S^3 \subset \mathbb{R}^4$ with $q \in \mathbb{H}_1 \sim S^3 \subset \mathbb{R}^4$.
• This segment is inside of the unit sphere in $\mathbb{R}^4$, and does not represent any connecting curve on the unit sphere $S^3 \sim \mathbb{H}_1$.
• Therefore the rotation is determined by normalizing the found quaternion.

This simple algorithm is almost perfect. The only problem is that after finding the interpolation points on the secant between the given points and normalizing the found quaternions, the resulted unit quaternions are not equidistant quaternions. This problem is solved by the following algorithm.

2. Slerp (Spherical linear interpolation)

Let $p, q \in \mathbb{H}_1,\ t \in [0,\, 1]$ and φ $(0 < \varphi < \pi)$ the angle between p and q. Then

$$\mathrm{Slerp}(p,\, q,\, t) = p(\overline{p}\,q)^t = p^{1-t}q^t = p\left[\frac{\sin((1-t)\varphi)}{\sin\varphi}\right] + q\left[\frac{\sin(t\,\varphi)}{\sin\varphi}\right]. \tag{4.144}$$

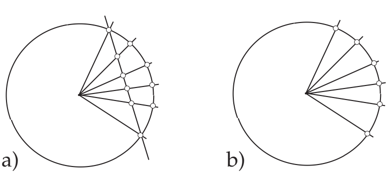


Figure 4.5

• Interpolation along the great circle on the unit sphere $S^3 \subset \mathbb{R}^4$, p and q are connected;
• The shortest connection is chosen, $-\mathrm{Sc}\,(p\,q) = \langle p,\, q\rangle > 0$ must hold (where $\langle\,,\,\rangle$ denotes the dot product of p and q in $\mathbb{R}^4$).

In **Fig.4.5** the interpolations according to Lerp(a) and Slerp(b) are compared.

Special case $p = 1$: Let $p = 1 = (1,\, 0,\, 0,\, 0)^{\mathrm{T}}$ and $q = \cos\varphi + \underline{\mathbf{n}}_{\mathbf{q}}\sin\varphi$, then

$$\mathrm{Slerp}(p,\, q,\, t) = \cos(t\,\varphi) + \underline{\mathbf{n}}_{\mathbf{q}}\sin(t\,\varphi)\,. \tag{4.145}$$

Special case equidistant grids: Let $\psi = \dfrac{\varphi}{n}$, then

$$q_k := \text{Slerp}(p,\, q,\, \frac{k}{n}) = \frac{1}{\sin\varphi}\left(\sin(\varphi - k\,\psi)\,p + \sin(k\,\psi)q\right), \quad k = 0,\, 1,\, \ldots,\, n\,. \tag{4.146}$$

Interpretation of the Slerp interpolation: To show the equivalence of the two expressions in (4.144) first $Q = p^{-1}q = \dfrac{\overline{p}}{|p|^2}q = \overline{p}\,q$ is calculated. Since $p, q \in \mathbb{H}_1$ the scalar part is

$$Q_0 = \text{Sc}\,Q = \text{Sc}\,(\overline{p}\,q) = \langle p,\, q\rangle = \cos\varphi\,. \tag{4.147}$$

Since $p = p\cdot 1$, and $q = p\,p^{-1}q = p\,Q$, the interpolation between 1 and Q is multiplied by p to keep the interpolation between p and q.

$$\begin{aligned} Q(t) &= \frac{\sin((1-t)\varphi)}{\sin\varphi} + Q\frac{\sin(t\varphi)}{\sin\varphi} = \frac{\sin((1-t)\varphi)}{\sin\varphi} + \cos\varphi\frac{\sin(t\varphi)}{\sin\varphi} + \vec{n}_{\vec{Q}}\frac{\sin(t\varphi)}{\sin\varphi} \\ &= \frac{\sin\varphi\cos(t\varphi) - \sin(t\varphi)\cos\varphi + \sin(t\varphi)\cos\varphi}{\sin\varphi} + \vec{n}_{\vec{Q}}\frac{\sin(t\varphi)\sin\varphi}{\sin\varphi} \\ &= \cos(t\varphi) + \vec{n}_{\vec{Q}}\sin(t\varphi) = e^{t\,\vec{n}_{\vec{Q}}\varphi} = e^{t\log Q} = Q^t\,. \end{aligned} \tag{4.148}$$

It follows that

$$q(t) = p\,Q(t) = p\frac{\sin((1-t)\varphi)}{\sin\varphi} + q\frac{\sin(t\varphi)}{\sin\varphi} = p\,Q^t = p(p^{-1}q)^t = p^{1-t}q^t\,. \tag{4.149}$$

3. Squad (spherical and quadrangle)

For $q_i,\, q_{i+1} \in \mathbb{H}_1$ and $t \in [0, 1]$ the rule is

$$\text{Squad}(q_i\; q_{i+1},\, s_i,\, s_{i+1},\, t) = \text{Slerp}(\text{Slerp}(q_i,\, q_{i+1},\, t), \text{Slerp}(s_i,\, s_{i+1},\, t),\, 2t(1-t)) \tag{4.150}$$

$$\text{with} \quad s_i = q_i \exp\left(-\frac{\log(q_i^{-1}q_{i+1}) + \log(q_i^{-1}q_{i-1})}{4}\right)\,.$$

- The resulted curve is similar to a Bézier curve, but it keeps the spherical instead of the linear interpolation.
- The algorithm produces an interpolation curve for a sequence of quaternions $q_0,\, q_1,\, \ldots,\, q_N$.
- The expression is not defined in the first and last interval, since q_{-1} is necessary to calculate s_0 and q_{N+1} to calculate s_N. A possible way out is to choose $s_0 = q_0$ and $s_N = q_N$, (or to define q_{-1} and q_{N+1}).

There are additional algorithms based on quaternions: nlerp, log-lerp, islerp, quaternion de Casteljau-splines.

4.4.3.2 Interpolation by Rotation matrices

The Slerp-algorithm can be described completely analogously with the help of rotation matrices. The logarithm of a 3×3 rotation matrix $\mathbf{R}$ is needed (i.e. an element of group $SO(3)$) and it is defined by a group-theoretical context as the skew-symmetric matrix $\mathbf{r}$ (i.e. an element of the Lie group so(3)), for which $e^{\mathbf{r}} = \mathbf{R}$. Then the Slerp-algorithm can be used to interpolate between rotation matrices $\mathbf{R}_0$ and $\mathbf{R}_1$,which is described as

$$\mathbf{R}(t) = \mathbf{R}_0(\mathbf{R}_0^{-1}\mathbf{R}_1)^t = \mathbf{R}_0\exp(t\log(\mathbf{R}_0^{-1}\mathbf{R}_1))\,. \tag{4.151}$$

In general it is more simple to use the quaternions based algorithm and to determine $\mathbf{R}(t)$ from $q(t)$ according to the calculations of the rotation matrix representing the unit quaternion.

4.4.3.3 Stereographic Projection

If $1 \in \mathbb{H}_1 \sim S^3$ is taken as the North pole of the three dimensional sphere S^3, then unit quaternions or elements of the three dimensional sphere can be mapped by the stereographic projection $\mathbb{H}_1 \ni q \mapsto$

$(1+q)(1-q)^{-1} \in \mathbb{H}_0 \sim \mathbb{R}^3$ into pure quaternions or into $\mathbb{R}^3$ respectively. The corresponding inverse mapping is

$$\mathbb{R}^3 \sim \mathbb{H}_0 \ni p \mapsto (p-1)(p+1)^{-1} \in \mathbb{H}_1 \sim S^3 . \tag{4.152}$$

4.4.3.4 Satellite navigation

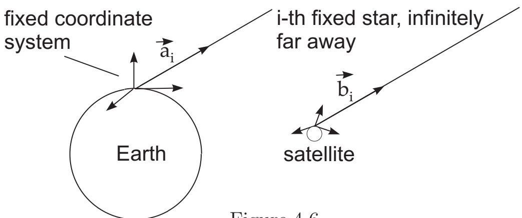


Figure 4.6

The orientation of an artificial satellite circulating around the Earth is to be determined. The fixed stars are considered to be infinitely far away, so their direction with respect to the Earth and the satellite are identical (see **Fig.4.6**). Any difference in measurements can be deduced from the different coordinate systems and so from the relative rotation of the coordinate systems.

Let $\vec{\mathbf{a}}_i$ be the unit vector pointing into the direction of the i-th fixed star from in the Earth's fixed coordinate system, and $\vec{\mathbf{b}}_i$ be the unit vector pointing into the direction of the i-th fixed star in the satellite's fixed coordinate system. The relative rotation of both coordinate systems can be described by a unit quaternion $h \in \mathbb{H}_1$:

$$\vec{\mathbf{b}}_i = h\,\vec{\mathbf{a}}_i\,\overline{h} . \tag{4.153}$$

If more fixed stars are considered, and the data are overlapped by measuring errors, then the solution is determined by the least squares method, i.e. as the minimum of (4.154), where h is a unit quaternion and $\underline{\mathbf{a}}_i = \vec{\mathbf{a}}_i$ and $\underline{\mathbf{b}}_i = \vec{\mathbf{b}}_i$ are unit vectors:

$$\begin{aligned} Q^2 &= \sum_{i=1}^{n} |\vec{\mathbf{b}}_i - h\,\vec{\mathbf{a}}_i\,\overline{h}|^2 = \sum_{i=1}^{n} (\vec{\mathbf{b}}_i - h\,\vec{\mathbf{a}}_i\,\overline{h}) \cdot (\vec{\mathbf{b}}_i - h\,\vec{\mathbf{a}}_i\,\vec{\mathbf{h}}) \\ &= \sum_{i=1}^{n} (\underline{\mathbf{b}}_i - h\,\underline{\mathbf{a}}_i\,\overline{h})\overline{(\underline{\mathbf{b}}_i - h\,\underline{\mathbf{a}}_i\,\overline{h})} = \sum_{i=1}^{n} (2 - \underline{\mathbf{b}}_i\,h\,\overline{\underline{\mathbf{a}}}_i\,\overline{h} - h\,\underline{\mathbf{a}}_i\,\overline{h}\,\overline{\underline{\mathbf{b}}}_i) . \end{aligned} \tag{4.154}$$

Since the group $\mathbb{H}_1$ of the unit quaternions form a Lie group, the critical points of Q^2 can be determined by the help of derivative

$$\partial_v h = \lim_{\vartheta \to 0} \frac{e^{\vartheta\, v} h - h}{\vartheta} = vh \quad (v, h \text{ quaternions}, \vartheta \text{ real}) \tag{4.155}$$

from

$$\partial_v Q^2 = -\sum_{i=1}^{n} \left(\underline{\mathbf{b}}_i\,\underline{\mathbf{v}}\,h\,\overline{\underline{\mathbf{a}}}_i\,\overline{h} + \underline{\mathbf{b}}_i\,h\,\overline{\underline{\mathbf{a}}}_i\,\overline{(\underline{\mathbf{v}}h)} + \underline{\mathbf{v}}\,h\,\underline{\mathbf{a}}_i\,\overline{h}\,\overline{\underline{\mathbf{b}}}_i + h\,\underline{\mathbf{a}}_i \overline{(\underline{\mathbf{v}}\,h)}\,\overline{\underline{\mathbf{b}}}_i \right) = 0 . \tag{4.156}$$

Here $\underline{\mathbf{v}}$, $\underline{\mathbf{b}}_i$ and $h\,\underline{\mathbf{a}}_i\,\overline{h}$ are pure quaternions, so $\overline{\underline{\mathbf{v}}} = -\underline{\mathbf{v}}$, and therefore (4.156) can be simplified:

$$\partial_{\underline{\mathbf{v}}} Q^2 = -4\underline{\mathbf{v}} \cdot \left(\sum_{i=1}^{n} h\,\vec{\mathbf{a}}_i\,\overline{h} \times \underline{\mathbf{b}}_i \right) = 0 . \tag{4.157}$$

Since here $\underline{\mathbf{v}}$ is arbitrary, this expression vanishes if

$$\sum_{i=1}^{n} h\,\vec{\mathbf{a}}_i\,\overline{h} \times \underline{\mathbf{b}}_i = \underline{\mathbf{0}} . \tag{4.158}$$

Let $\mathbf{R}$ be the rotation matrix represented by the unit quaternion h, i.e. $h\,\vec{\mathbf{a}}_i \overline{h} = \mathbf{R}\,\vec{\mathbf{a}}_i$. With the 3×3

matrix

$$\mathbf{K}(\vec{\mathbf{a}}) = \begin{pmatrix} 0 & -a_z & a_y \\ a_z & 0 & -a_x \\ -a_y & a_x & 0 \end{pmatrix} \tag{4.159}$$

defined by vector $\vec{\mathbf{a}} = (a_x,\, a_y,\, a_z) \in \mathbb{R}^3$ for any vector $\vec{\mathbf{b}} \in \mathbb{R}^3$ one gets:

$$\mathbf{K}(\vec{\mathbf{a}})\,\vec{\mathbf{b}} = \vec{\mathbf{a}} \times \vec{\mathbf{b}}, \qquad \mathbf{K}(\mathbf{K}(\vec{\mathbf{a}})\,\vec{\mathbf{b}}) = \vec{\mathbf{b}}\,\vec{\mathbf{a}}^{\mathrm{T}} - \vec{\mathbf{a}}\,\vec{\mathbf{b}}^{\mathrm{T}}. \tag{4.160}$$

From this relation the critical points of the minimum problems are determined:

$$\sum_{i=1}^{n} \mathbf{K}\,(\mathbf{R}\,\vec{\mathbf{a}}_i \times \vec{\mathbf{b}}_i) = \mathbb{O} \iff \sum_{i=1}^{n} (\vec{\mathbf{b}}_i\,\vec{\mathbf{a}}_i^{\mathrm{T}}\,\mathbf{R}^{\mathrm{T}} - \mathbf{R}\,\vec{\mathbf{a}}_i\,\vec{\mathbf{b}}_i^{\mathrm{T}}) = \mathbb{O} \iff \mathbf{R}\,\mathbf{P} = \mathbf{P}^{\mathrm{T}}\mathbf{R}^{\mathrm{T}} \tag{4.161}$$

where $\mathbf{P} = \sum_{i=1}^{n} \vec{\mathbf{a}}_i\,\vec{\mathbf{b}}_i^{\mathrm{T}}$. If $\mathbf{P}$ is decomposed into the product $\mathbf{P} = \mathbf{R}_p^{\mathrm{T}}\mathbf{S}$, where matrix S is symmetric and $\mathbf{P} = \mathbf{R}_p$ is a rotation matrix, then from (4.161) follows

$$\mathbf{R}\mathbf{R}_p^{\mathrm{T}}\mathbf{S} = \mathbf{S}\mathbf{R}_p\mathbf{R}^{\mathrm{T}}\,, \tag{4.162}$$

and

$$\mathbf{R} = \mathbf{R}_p \tag{4.163}$$

is obviously a solution, since in this case $\mathbf{R}_p\mathbf{R}_p^{\mathrm{T}}\mathbf{S} = \mathbf{S} = \mathbf{S}\mathbf{R}_p\mathbf{R}_p^{\mathrm{T}}$, because $\mathbf{R}_p\mathbf{R}_p^{\mathrm{T}} = \mathbf{E}$. However there are three other solutions, namely

$$\mathbf{R} = \mathbf{R}_j\mathbf{R}_p \quad (j = 1, 2, 3)\,, \tag{4.164}$$

where $\mathbf{R}_j$ denotes the rotation by π around the j-th eigenvector of $\mathbf{S}$, i.e., there is $\mathbf{R_j}\mathbf{S}\mathbf{R_j} = \mathbf{S}$. That $\mathbf{R} = \mathbf{R}_j\mathbf{R}_\nu$ is a solution of (4.162) which can be seen from $\mathbf{R}_j\mathbf{R}_p\mathbf{R}_p^{\mathrm{T}}\mathbf{S} = \mathbf{S}\mathbf{R}_p\mathbf{R}_p^{\mathrm{T}}\mathbf{R}_j^{\mathrm{T}} \iff \mathbf{R}_j\mathbf{S} = \mathbf{S}\mathbf{R}_j^{\mathrm{T}} \iff \mathbf{R}_j\mathbf{S}\mathbf{R}_j = \mathbf{S}$.
The solution for which Q^2 is minimal is

$$\mathbf{R} = \begin{cases} \mathbf{R}_p\,, & \text{falls} \quad \det \mathbf{P} > 0\,, \\ \mathbf{R}_{j_0}\,\mathbf{R}_p\,, & \text{falls} \quad \det \mathbf{P} < 0\,, \end{cases} \tag{4.165}$$

where $\mathbf{R}_{j_0}$ is the rotation by π around the eigenvector of $\mathbf{S}$ associated with the eigenvalue of the smallest absolute value.

4.4.3.5 Vector Analysis

If the ∇ operator (see (13.67), 13.2.6.1, p. 715) and a vector $\vec{\mathbf{v}}$ (see 13.1.3, p. 704) are identified with ∇_Q and $\underline{\mathbf{v}}$ in quaternion calculus, i.e.

$$\nabla_Q = \mathbf{i}\frac{\partial}{\partial x} + \mathbf{j}\frac{\partial}{\partial y} + \mathbf{k}\frac{\partial}{\partial z}\,, \tag{4.166}$$

$$\underline{\mathbf{v}}(x,\, y,\, z) = v_1(x,\, y,\, z)\mathbf{i} + v_2(x,\, y,\, z)\mathbf{j} + v_3(x,\, y,\, z)\mathbf{k} \tag{4.167}$$

with $\mathbf{i}\,,\mathbf{j}\,,\mathbf{k}$ (according to (4.107), p. 290), then the multiplication rule for quaternions (see (4.109b), p. 291) gives

$$\nabla_Q\underline{\mathbf{v}} = -\nabla \cdot \vec{\mathbf{v}} + \nabla \times \vec{\mathbf{v}} = -\mathrm{div}\,\vec{\mathbf{v}} + \mathrm{rot}\,\vec{\mathbf{v}}, \tag{4.168}$$

(see also ■ in 4.4.1.1, **4.** p. 291).
Substituting

$$D = \frac{\partial}{\partial t} + \mathbf{i}\frac{\partial}{\partial x} + \mathbf{j}\frac{\partial}{\partial y} + \mathbf{k}\frac{\partial}{\partial z} \quad \text{and} \tag{4.169a}$$

$$\begin{aligned} w(t,\, x,\, y,\, z) &= w_0(t,\, x,\, y,\, z) + w_1(t,\, x,\, y,\, z)\mathbf{i} + w_2(t,\, x,\, y,\, z)\mathbf{j} + w_3(t,\, x,\, y,\, z)\mathbf{k} \\ &= w_0(t,\, x,\, y,\, z) + \underline{\mathbf{w}}(t,\, x,\, y,\, z), \end{aligned} \tag{4.169b}$$

then

$$D\,w = \frac{\partial}{\partial t}\,w_0 - \operatorname{div}\underline{\mathbf{w}} + \operatorname{rot}\,\underline{\mathbf{w}} + \operatorname{grad}\,w_0\,. \qquad (4.169c)$$

Especially, for an arbitrary twice continuously differentiable function $f(t,x,y,z)$

$$\nabla_Q\overline{\nabla}_Q\,f = \overline{\nabla}_Q\nabla_Q\,f = \nabla\nabla\,f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2} = \Delta_3\,f \quad \text{und} \qquad (4.170a)$$

$$\nabla_Q\nabla_Q\,f = -\nabla\nabla\,f = -\frac{\partial^2 f}{\partial x^2} - \frac{\partial^2 f}{\partial y^2} - \frac{\partial^2 f}{\partial z^2} = -\Delta_3\,f\,, \qquad (4.170b)$$

where Δ_3 denotes the Laplace operator in $\mathbb{R}^3$ (see (13.75) in 13.2.6.5, p. 716).

$$D\overline{D}f = \overline{D}Df = \frac{\partial^2 f}{\partial t^2} + \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2} = \Delta_4 f \qquad (4.170c)$$

where Δ_4 denotes the Laplace-operator in $\mathbb{R}^4$. ∇_Q, just as D is often called the Dirac- or Cauchy-Riemann operator.

4.4.3.6 Normalized Quaternions and Rigid Body Motion

1. Biquaternions

A biquaternion $\check{h}$ has the form

$$\check{h} = h_0 + \epsilon\,h_1, \quad \text{with} \quad h_0,\, h_1 \in \mathbb{H} \qquad (4.171)$$

Here ϵ is the dual unit, that commutes with every quaternion, furthermore $\epsilon^2 = 0$. The multiplication is the usual quaternion multiplication (see 4.115, p. 292).

2. Rigid Body Motion

By the help of unit biquaternions, i.e. biquaternions with

$$\check{h}\overline{\check{h}} = (h_0 + \epsilon h_1)(\overline{h}_0 + \epsilon\overline{h}_1) = 1 \iff \begin{cases} h_0\,\overline{h}_0 = 1\,, \\ h_0\,\overline{h}_1 + h_1\overline{h}_0 = 0\,, \end{cases} \qquad (4.172)$$

rigid-body motion (rotation and translation after each other) can be described in $\mathbb{R}^3$.

Table 4.1 Rigid body motions with biquaternions

Element	Representation by
point $\underline{\mathbf{p}} = (p_x,\, p_y,\, p_z)$ in space	$\check{p} = 1 + \underline{\mathbf{p}}\,\epsilon$ with $\underline{\mathbf{p}} = p_x\,\mathbf{i} + p_y\,\mathbf{j} + p_z\,\mathbf{k}$
rotations	$r \in \mathbb{H}_1$ unit quaternions
translations $\underline{\mathbf{t}} = (t_x,\, t_y,\, t_z)$	$1 + \frac{1}{2}\underline{\mathbf{t}}\,\epsilon$ with $\underline{\mathbf{t}} = t_x\,\mathbf{i} + t_y\,\mathbf{j} + t_z\,\mathbf{k}$

The unit biquaternions

$$\check{h} = h_0 + h_1\,\epsilon = \left(1 + \frac{1}{2}\,\underline{\mathbf{t}}\,\epsilon\right)\,r = r + \frac{1}{2}\underline{\mathbf{t}}\,r\,\epsilon, \quad \underline{\mathbf{t}} \in \mathbb{H}_0\,, \quad r \in \mathbb{H}_1\,, \qquad (4.173)$$

give a double covering over the group $\mathbf{SE(3)}$ of the rigid-body motion in $\mathbb{R}^3$ since $\check{h}$ and $-\check{h}$ describe the same rigid-body motion.

4.5 Systems of Linear Equations

4.5.1 Linear Systems, Pivoting

4.5.1.1 Linear Systems

A *linear system* contains m linear forms

$$\begin{array}{l} y_1 = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n + a_1 \\ y_2 = a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n + a_2 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ y_m = a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n + a_m \end{array} \quad \text{or} \quad \underline{\mathbf{y}} = \mathbf{A}\underline{\mathbf{x}} + \underline{\mathbf{a}} \tag{4.174a}$$

with

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & & & \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}, \quad \underline{\mathbf{a}} = \begin{pmatrix} a_1, \\ a_2, \\ \vdots \\ a_m \end{pmatrix}, \quad \underline{\mathbf{x}} = \begin{pmatrix} x_1, \\ x_2, \\ \vdots \\ x_n \end{pmatrix}, \quad \underline{\mathbf{y}} = \begin{pmatrix} y_1, \\ y_2, \\ \vdots \\ y_m \end{pmatrix}. \tag{4.174b}$$

The elements $a_{\mu\nu}$ of the matrix $\mathbf{A}$ of size (m, n) and the components a_μ $(\mu = 1, 2, \ldots, m)$ of the column vector $\underline{\mathbf{a}}$ are constants. The components x_ν $(m = 1, 2, \ldots, n)$ of the column vector $\underline{\mathbf{x}}$ are the *independent variables*, the components y_μ $(\mu = 1, 2, \ldots, m)$ of the column vector $\underline{\mathbf{y}}$ are the *dependent variables*.

4.5.1.2 Pivoting

1. Scheme of Pivoting

If an element a_{ik} is not equal to zero in (4.174a), the variable y_i can be exchanged for an independent one and the variable x_k for a dependent one in a so-called *pivoting step*. The pivoting step is the basic element of pivoting, the method by which for instance systems of linear equations and linear optimization problems can be solved. The pivoting step is achieved by the schemes

$$\begin{array}{c|ccccccc} & x_1 & x_2 & \cdots & x_k & \cdots & x_n & 1 \\ \hline y_1 & a_{11} & a_{12} & \cdots & a_{1k} & \cdots & a_{1n} & a_1 \\ y_2 & a_{21} & a_{22} & \cdots & a_{2k} & \cdots & a_{2n} & a_2 \\ \vdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ y_i & a_{i1} & a_{i2} & \cdots & \boxed{a_{ik}} & \cdots & a_{in} & a_i \\ \vdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ y_m & a_{m1} & a_{m2} & \cdots & a_{mk} & \cdots & a_{mn} & a_m \\ \hline x_k & \alpha_{i1} & \alpha_{i2} & \cdots & \alpha_{ik} & \cdots & \alpha_{in} & \alpha_i \end{array}, \qquad \begin{array}{c|ccccccc} & x_1 & x_2 & \cdots & y_i & \cdots & x_n & 1 \\ \hline y_1 & \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1k} & \cdots & \alpha_{1n} & \alpha_1 \\ y_2 & \alpha_{21} & \alpha_{22} & \cdots & \alpha_{2k} & \cdots & \alpha_{2n} & \alpha_2 \\ \vdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ x_k & \alpha_{i1} & \alpha_{i2} & \cdots & \alpha_{ik} & \cdots & \alpha_{in} & \alpha_i \\ \vdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ y_m & \alpha_{m1} & \alpha_{m2} & \cdots & \alpha_{mk} & \cdots & \alpha_{mn} & \alpha_m \end{array}, \tag{4.175}$$

where the left scheme corresponds to the system (4.174a).

2. Pivoting Rules

The framed element in the scheme a_{ik} $(a_{ik} \neq 0)$ is called the *pivot element*; it is at the intersection of the *pivot column* and *pivot row*. The elements $\alpha_{\mu\nu}$ and α_μ of the new scheme on the right-hand side will be calculated by the following *pivoting rules*:

1. $$\alpha_{ik} = \frac{1}{a_{ik}}, \tag{4.176a}$$ **2.** $$\alpha_{\mu k} = \frac{a_{\mu k}}{a_{ik}} \quad (\mu = 1, \ldots, m; \mu \neq i), \tag{4.176b}$$

3. $$\alpha_{i\nu} = -\frac{a_{i\nu}}{a_{ik}}, \quad \alpha_i = -\frac{a_i}{a_{ik}} \quad (\nu = 1, 2, \ldots, n; \nu \neq k), \tag{4.176c}$$

4. $$\alpha_{\mu\nu} = a_{\mu\nu} - a_{\mu k}\frac{a_{i\nu}}{a_{ik}} = a_{\mu\nu} + a_{\mu k}\alpha_{i\nu}, \quad \alpha_\mu = a_\mu + a_{\mu k}\alpha_i$$

$$(\text{for every } \mu \neq i \text{ and every } \nu \neq k). \tag{4.176d}$$

To make the calculations easier (rule 4) one writes the elements $\alpha_{i\nu}$ in the $(m+1)$-th row of the pivoting scheme (cellar row). With this pivoting rule one can change further variables.

4.5.1.3 Linear Dependence

The linear forms (4.174a) are linearly independent (see 9.1.2.3, **2.**, p. 553), if all y_μ can be changed for an independent variable x_ν . The linear independence will be used, for instance to determine the rank of a matrix. Otherwise, the dependence relation can be found directly from the scheme.

■

	x_1	x_2	x_3	x_4	1
y_1	2	1	1	0	−2
y_2	1	−1	0	0	2
y_3	1	5	2	0	0
y_4	0	2	0	1	0

After three pivoting steps (for instance $y_4 \to x_4$, $y_2 \to x_1$, $y_1 \to x_3$) the table becomes:

	y_2	x_2	y_1	y_4	1
x_3	−2	−3	1	0	6
x_1	1	1	0	0	−2
y_3	−3	$\boxed{0}$	2	0	10
x_4	0	−2	0	1	0

No further change is possible because $\alpha_{32} = 0$, and one can see the dependence relation $y_3 = 2y_1 - 3y_2 + 10$. Also for another sequence of pivoting, there remains one pair of not exchangeable variables.

4.5.1.4 Calculation of the Inverse of a Matrix

If $\mathbf{A}$ is a regular matrix of size (n, n), then the inverse matrix $\mathbf{A}^{-1}$ can be obtained after n steps using the pivoting procedure for the system $\underline{\mathbf{y}} = \mathbf{A}\underline{\mathbf{x}}$.

■ $\mathbf{A} = \begin{pmatrix} 3 & 5 & 1 \\ 2 & 4 & 5 \\ 1 & 2 & 2 \end{pmatrix} \Longrightarrow$

	x_1	x_2	x_3
y_1	3	5	1
y_2	2	4	5
y_3	$\boxed{1}$	2	2

,

	y_3	x_2	x_3
y_1	3	−1	−5
y_2	2	0	$\boxed{1}$
x_1	1	−2	−2

,

	y_3	x_2	y_2
y_1	13	$\boxed{-1}$	−5
x_3	−2	0	1
x_1	5	−2	−2

,

	y_3	y_1	y_2
x_2	13	−1	−5
x_3	−2	0	1
x_1	−21	2	8

.

After rearranging the elements one gets $\mathbf{A}^{-1} = \begin{pmatrix} 2 & 8 & -21 \\ -1 & -5 & 13 \\ 0 & 1 & -2 \end{pmatrix}$. (The columns are to be arranged with respect to the indices of y_i, the rows with respect to the indices of x_k.)

4.5.2 Solution of Systems of Linear Equations

4.5.2.1 Definition and Solvability

1. System of Linear Equations

A system of m linear equations with n unknowns $x_1, x_2, \ldots, x_n$

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= a_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= a_2 \\ &\vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= a_m \end{aligned} \quad \text{or briefly} \quad \mathbf{A}\underline{\mathbf{x}} = \underline{\mathbf{a}}, \tag{4.177a}$$

is called a *linear equation system*. Here the following designations are used:

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & & & \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}, \qquad \underline{\mathbf{a}} = \begin{pmatrix} a_1, \\ a_2, \\ \vdots \\ a_m \end{pmatrix}, \qquad \underline{\mathbf{x}} = \begin{pmatrix} x_1, \\ x_2, \\ \vdots \\ x_n \end{pmatrix}. \tag{4.177b}$$

If the column vector $\underline{\mathbf{a}}$ is the zero vector ($\underline{\mathbf{a}} = \underline{\mathbf{0}}$), then the system of equations is called a *homogeneous system*, otherwise ($\underline{\mathbf{a}} \neq \underline{\mathbf{0}}$) it is called an *inhomogeneous system of equations*. The coefficients $a_{\mu\nu}$ of the system are the elements of the so-called *matrix of coefficients* $\mathbf{A}$, and the components a_μ of the column vector $\underline{\mathbf{a}}$ are the *constant terms (absolute terms)*.

2. Solvability of a Linear System of Equations

A linear system of equations is called *solvable* or *consistent* or *compatible* if it has a solution, i.e., there exists at least one vector $\underline{\mathbf{x}} = \underline{\boldsymbol{\alpha}}$ such that (4.177a) is an identity. Otherwise, it is called inconsistent. The existence and uniqueness of the solution depend on the rank of the *augmented matrix* $(\mathbf{A}, \underline{\mathbf{a}})$. One gets the augmented matrix by attaching the vector $\underline{\mathbf{a}}$ to the matrix $\mathbf{A}$ as its $(n+1)$-th column.

1. General Rules for Inhomogeneous Linear Systems of Equations An inhomogeneous linear system of equations $\mathbf{A}\underline{\mathbf{x}} = \underline{\mathbf{a}}$ has at least one solution if

$$\operatorname{rank}(\mathbf{A}) = \operatorname{rank}(\mathbf{A}, \underline{\mathbf{a}}), \tag{4.178a}$$

is valid. Furthermore, if r denotes the rank of $\mathbf{A}$, i.e., $r = \operatorname{rank}(\mathbf{A})$, then

a) for $r = n$ the system has a unique solution, (4.178b)

b) for $r < n$ the system has infinitely many solutions, (4.178c)

i.e., the values of $n - r$ unknowns as parameters can be chosen freely.

■ **A:**

$$\begin{aligned} x_1 - 2x_2 + 3x_3 - x_4 + 2x_5 &= 2 \\ 3x_1 - x_2 + 5x_3 - 3x_4 - x_5 &= 6 \\ 2x_1 + x_2 + 2x_3 - 2x_4 - 3x_5 &= 8 \end{aligned}$$

The rank of $\mathbf{A}$ is 2, the rank of the *augmented matrix of coefficients* $(\mathbf{A}, \underline{\mathbf{a}})$ is 3, i.e., the system is inconsistent.

■ **B:**

$$\begin{aligned} x_1 - x_2 + 2x_3 &= 1 \\ x_1 - 2x_2 - x_3 &= 2 \\ 3x_1 - x_2 + 5x_3 &= 3 \\ -2x_1 + 2x_2 + 3x_3 &= -4 \end{aligned}$$

Both the matrices $\mathbf{A}$ and $(\mathbf{A}, \underline{\mathbf{a}})$ have rank equal to 3. Because $r = n = 3$ the system has a unique solution: $x_1 = \frac{10}{7}, \quad x_2 = -\frac{1}{7}, \quad x_3 = -\frac{2}{7}$.

■ **C:**

$$\begin{aligned} x_1 - x_2 + x_3 - x_4 &= 1 \\ x_1 - x_2 - x_3 + x_4 &= 0 \\ x_1 - x_2 - 2x_3 + 2x_4 &= -\frac{1}{2} \end{aligned}$$

Both the matrices $\mathbf{A}$ and $(\mathbf{A}, \underline{\mathbf{a}})$ have rank equal to 2. The system is consistent but because $r < n$ it does not have a unique solution. Therefore $n - r = 2$ unknowns can be considered as free parameters: $x_2 = x_1 - \frac{1}{2}, \quad x_3 = x_4 + \frac{1}{2}, \quad (x_1, x_4$ arbitrary values).

■ **D:**

$$\begin{aligned} x_1 + 2x_2 - x_3 + x_4 &= 1 \\ 2x_1 - x_2 + 2x_3 + 2x_4 &= 2 \\ 3x_1 + x_2 + x_3 + 3x_4 &= 3 \\ x_1 - 3x_2 + 3x_3 + x_4 &= 0 \end{aligned}$$

There is the same number of equations as unknowns but the system has no solution because $\operatorname{rank}(\mathbf{A}) = 2$, and $\operatorname{rank}(\mathbf{A}, \underline{\mathbf{a}}) = 3$.

2. Trivial Solution and Fundamental System of Homogeneous Systems

a) The homogeneous system of equations $\mathbf{A}\underline{\mathbf{x}} = \underline{\mathbf{0}}$ always has a solution, the so-called trivial solution

$$x_1 = x_2 = \ldots = x_n = 0. \tag{4.179a}$$

(The equality $\operatorname{rank}(\mathbf{A}) = \operatorname{rank}(\mathbf{A}, \underline{\mathbf{0}})$ always holds.)

b) If the homogeneous system has the non-trivial solutions $\underline{\boldsymbol{\alpha}} = (\alpha_1, \alpha_2, \ldots, \alpha_n)$ and $\underline{\boldsymbol{\beta}} = (\beta_1, \beta_2, \ldots, \beta_n)$, i.e., $\underline{\boldsymbol{\alpha}} \neq \underline{\mathbf{0}}$ and $\underline{\boldsymbol{\beta}} \neq \underline{\mathbf{0}}$, then $\underline{\mathbf{x}} = s\,\underline{\boldsymbol{\alpha}} + t\,\underline{\boldsymbol{\beta}}$ is also a solution with arbitrary constants s and t, i.e., any linear combination of the solutions is a solution as well.

Suppose, the system has exactly l non-trivial linearly independent solutions $\underline{\boldsymbol{\alpha}}_1, \underline{\boldsymbol{\alpha}}_2, \ldots, \underline{\boldsymbol{\alpha}}_l$. Then these solutions form a so-called *fundamental system* of solutions (see 9.1.2.3, **2.**, p. 553), and the general solution of the homogeneous system of equations has the form

$$\underline{\mathbf{x}} = k_1\underline{\boldsymbol{\alpha}}_1 + k_2\underline{\boldsymbol{\alpha}}_2 + \cdots + k_l\underline{\boldsymbol{\alpha}}_l \qquad (k_1, k_2, \ldots, k_l \text{ arbitrary constants}). \tag{4.179b}$$

If the rank r of the coefficient matrix $\mathbf{A}$ of the homogeneous system of equations is less than the number of unknowns n, i.e., $r < n$, then the system of equations has $l = n - r$ linearly independent non-trivial solutions. If $r = n$, then the solution is unique, i.e., the homogeneous system has only the trivial

solution.

To determine a fundamental system in the case $r < n$ one chooses $n - r$ unknowns as free parameters, and expresses the remaining unknowns in terms of them. If reordering the equations and the unknowns so that the subdeterminant of order r in the left upper corner is not equal to zero, then one gets for instance:

$$\begin{aligned} x_1 &= x_1(x_{r+1}, x_{r+2}, \ldots, x_n) \\ x_2 &= x_2(x_{r+1}, x_{r+2}, \ldots, x_n) \\ \vdots\ &\ \vdots \qquad\qquad \vdots \\ x_r &= x_r(x_{r+1}, x_{r+2}, \ldots, x_n). \end{aligned} \tag{4.180}$$

Then one can get a fundamental system of solutions choosing the free parameters, for instance in the following way:

$$\begin{array}{lccccc} & x_{r+1} & x_{r+2} & x_{r+3} & \cdots & x_n \\ \hline \text{1. fundamental solution:} & 1 & 0 & 0 & \cdots & 0 \\ \text{2. fundamental solution:} & 0 & 1 & 0 & \cdots & 0 \\ \qquad\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ (n-r)\text{-th fundamental solution:} & 0 & 0 & 0 & \cdots & 1 \end{array}. \tag{4.181}$$

■ **E:**

$$\begin{aligned} x_1 - \ \ x_2 + 5x_3 - \ \ x_4 &= 0 \\ x_1 + \ \ x_2 - 2x_3 + 3x_4 &= 0 \\ 3x_1 - \ \ x_2 + 8x_3 + \ \ x_4 &= 0 \\ x_1 + 3x_2 - 9x_3 + 7x_4 &= 0 \end{aligned}$$

The rank of the matrix $\mathbf{A}$ is equal to 2. The system can be solved for x_1 and x_2 resulting to: $x_1 = -\frac{3}{2}x_3 - x_4$, $x_2 = \frac{7}{2}x_3 - 2x_4$ (x_3, x_4 arbitrary). Fundamental solutions are $\underline{\boldsymbol{\alpha}_1} = (-\frac{3}{2}, \frac{7}{2}, 1, 0)^{\mathrm{T}}$ and $\underline{\boldsymbol{\alpha}_2} = (-1, -2, 0, 1)^{\mathrm{T}}$.

4.5.2.2 Application of Pivoting

1. System of Linear Functions Corresponding to a Linear System of Equations

In order to solve (4.177a), a system of linear functions $\underline{\mathbf{y}} = \mathbf{A}\underline{\mathbf{x}} - \underline{\mathbf{a}}$ is assigned to the system of equations $\mathbf{A}\underline{\mathbf{x}} = \underline{\mathbf{a}}$ so the use of pivoting (see 4.5.1.2, p. 307) is possible:

$$\mathbf{A}\underline{\mathbf{x}} = \underline{\mathbf{a}} \quad (4.182\text{a}) \qquad \text{is equivalent to} \qquad \underline{\mathbf{y}} = \mathbf{A}\underline{\mathbf{x}} - \underline{\mathbf{a}} = \underline{\mathbf{0}}. \quad (4.182\text{b})$$

The matrix $\mathbf{A}$ is of size (m, n), $\underline{\mathbf{a}}$ is a column vector with m components, i.e., the number of equations m must not be equal to the number of unknowns n. After finishing the pivoting one substitutes $\underline{\mathbf{y}} = \underline{\mathbf{0}}$. The existence and uniqueness of the solution of $\mathbf{A}\underline{\mathbf{x}} = \underline{\mathbf{a}}$ can be seen directly from the last pivoting scheme.

2. Solvability of Linear Systems of Equations

The linear system of equations (4.182a) has a solution if one of the following two cases holds for the corresponding linear functions (4.182b):

Case 1: All y_μ ($\mu = 1, 2, \ldots, m$) can be exchanged for some x_ν. This means the corresponding system of linear functions is linearly independent.

Case 2: At least one y_σ cannot be exchanged for any x_ν, i.e.,

$$y_\sigma = \lambda_1 y_1 + \lambda_2 y_2 + \cdots + \lambda_m y_m + \lambda_0 \tag{4.183}$$

holds and also $\lambda_0 = 0$. This means the corresponding system of linear functions is linearly dependent.

3. Inconsistency of Linear Systems of Equations

The linear system of equations has no solution if in case 2 above $\lambda_0 \neq 0$ holds. In this case the system has contradictory equations.

$$\begin{aligned} x_1 - 2x_2 + 4x_3 - x_4 &= 2 \\ -3x_1 + 3x_2 - 3x_3 + 4x_4 &= 3 \\ 2x_1 - 3x_2 + 5x_3 - 3x_4 &= -1 \end{aligned}$$

■

	x_1	x_2	x_3	x_4	1
y_1	1	−2	4	−1	−2
y_2	−3	3	−3	4	−3
y_3	2	−3	5	−3	1

After three pivoting steps (for instance $y_1 \to x_1$, $y_3 \to x_4$, $y_2 \to x_2$) follows:

	y_1	y_2	x_3	y_3	1
x_1	$\frac{3}{2}$	$-\frac{3}{2}$	2	$-\frac{5}{2}$	1
x_2	$-\frac{1}{2}$	$-\frac{1}{2}$	3	$-\frac{1}{2}$	−2
x_4	$\frac{3}{2}$	$-\frac{1}{2}$	0	$-\frac{3}{2}$	3.

This calculation ends with case 1: y_1, y_2, y_3 and x_3 are independent variables. Substituting $y_1 = y_2 = y_3 = 0$, and $x_3 = t$ ($-\infty < t < \infty$ is a parameter) consequently, the solution is: $x_1 = 2t+1$, $x_2 = 3t-2$, $x_3 = t$, $x_4 = 3$.

4.5.2.3 Cramer's Rule

There is the very important special case when the number of equations is equal to the number of unknowns

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= a_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= a_2 \\ \vdots \qquad\qquad\qquad \vdots \quad &\;\; \vdots \;\; \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n &= a_n \end{aligned} \tag{4.184a}$$

and the determinant of the coefficients does not vanish, i.e.,

$$\mathrm{D} = \det \mathbf{A} \neq 0\,. \tag{4.184b}$$

In this case the unique solution of the system of equations (4.184a) can be given in an explicit and unique form:

$$x_1 = \frac{\mathrm{D}_1}{\mathrm{D}}, \quad x_2 = \frac{\mathrm{D}_2}{\mathrm{D}}, \quad \ldots, \quad x_n = \frac{\mathrm{D}_n}{\mathrm{D}}. \tag{4.184c}$$

D_ν denotes the determinant, which is obtained from D by replacing the elements $a_{\mu\nu}$ of the ν-th column of D by the constant terms a_μ, for instance

$$\mathrm{D}_2 = \begin{vmatrix} a_{11} & a_1 & a_{13} & \cdots & a_{1n} \\ a_{21} & a_2 & a_{23} & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_n & a_{n3} & \cdots & a_{nn} \end{vmatrix}. \tag{4.184d}$$

If $\mathrm{D} = 0$ and there is at least one $\mathrm{D}_\nu \neq 0$, then the system (4.184a) has no solution.
In the case $\mathrm{D} = 0$ and $\mathrm{D}_\nu = 0$ for all $\nu = 1, 2, \ldots, n$, then it is possible that the system has a solution but it is not unique. (see Remark p. 311).

■ $$\begin{aligned} 2x_1 + x_2 + 3x_3 &= 9 \\ x_1 - 2x_2 + x_3 &= -2 \\ 3x_1 + 2x_2 + 2x_3 &= 7. \end{aligned} \qquad \mathrm{D} = \begin{vmatrix} 2 & 1 & 3 \\ 1 & -2 & 1 \\ 3 & 2 & 2 \end{vmatrix} = 13,$$

$$\mathrm{D}_1 = \begin{vmatrix} 9 & 1 & 3 \\ -2 & -2 & 1 \\ 7 & 2 & 2 \end{vmatrix} = -13, \quad \mathrm{D}_2 = \begin{vmatrix} 2 & 9 & 3 \\ 1 & -2 & 1 \\ 3 & 7 & 2 \end{vmatrix} = 26, \quad \mathrm{D}_3 = \begin{vmatrix} 2 & 1 & 9 \\ 1 & -2 & -2 \\ 3 & 2 & 7 \end{vmatrix} = 39.$$

The system has the unique solution $x_1 = \dfrac{\mathrm{D}_1}{\mathrm{D}} = -1$, $x_2 = \dfrac{\mathrm{D}_2}{\mathrm{D}} = 2$, $x_3 = \dfrac{\mathrm{D}_3}{\mathrm{D}} = 3$.

Remark: From practical consideration the Cramer rule is not useful for higher-dimensional problems. As the dimension of the problem increases, the number of required operations increases very fast, so, for numerical solutions of linear systems of equations one uses the Gauss algorithm or pivoting or an

iteration procedure (see 19.1.1, p. 949).

4.5.2.4 Gauss's Algorithm

1. Gauss Elimination Method In order to solve the linear system of equations $\mathbf{A}\underline{\mathbf{x}} = \underline{\mathbf{a}}$ (4.177a) of m equations with n unknowns one can use the *Gauss elimination method*. With the help of an equation one unknown is to be eliminated from all the other equations. So one gets a system of $m-1$ equations and $n-1$ unknowns. This method will be repeated until the result is a system of equations in *row echelon form*, and from this form one can determine the existence and uniqueness of the solution easily, and the solution itself can be found if it exists.

2. Gauss Steps The first Gauss step is to be demonstrated on the augmented matrix of coefficients $(\mathbf{A}, \underline{\mathbf{a}})$ (see examples on p. 309):
Supposing $a_{11} \neq 0$, otherwise exchanging the first equation for another one. In the matrix

$$\left(\begin{array}{cccc|c} a_{11} & a_{12} & \cdots & a_{1n} & a_1 \\ a_{21} & a_{22} & \cdots & a_{2n} & a_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} & a_m \end{array}\right) \tag{4.185a}$$

the appropriate multiple of the first row is to be added to the others in order to make the coefficients of x_1 equal to zero, i.e., multiply the first row by $-\frac{a_{21}}{a_{11}}, -\frac{a_{31}}{a_{11}}, \ldots, -\frac{a_{m1}}{a_{11}}$ then add them to the second, third,..., m-th row. The transformed matrix has the form

$$\left(\begin{array}{cccc|c} a_{11} & a_{12} & \cdots & a_{1n} & a_1 \\ 0 & a'_{22} & \cdots & a'_{2n} & a'_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & a'_{m2} & \cdots & a'_{mn} & a'_m \end{array}\right). \tag{4.185b}$$

After applying this Gauss step $(r-1)$ times the result is a matrix in row echelon form

$$\left(\begin{array}{cccccccc|c} a_{11} & a_{12} & a_{13} & \ldots & & a_{1,r+1} & \ldots & a_{1n} & a_1 \\ 0 & a'_{22} & a'_{23} & \ldots & & a'_{2,r+1} & \ldots & a'_{2n} & a'_2 \\ 0 & 0 & a''_{33} & \ldots & & a''_{3,r+1} & \ldots & a''_{3n} & a''_3 \\ \ldots & \ldots & & \ldots & & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & & a^{(r-1)}_{r,r} & a^{(r-1)}_{r,r+1} & \ldots & a^{(r-1)}_{rn} & a^{(r-1)}_r \\ 0 & 0 & \ldots & \ldots & 0 & 0 & \ldots & 0 & a^{(r-1)}_{r+1} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \ldots & 0 & 0 & \ldots & 0 & a^{(r-1)}_m \end{array}\right). \tag{4.186}$$

3. Existence and Uniqueness of the Solution The Gauss steps are elementary row operations so they do not affect the rank of the matrix $(\mathbf{A}, \underline{\mathbf{a}})$, consequently the existence and uniqueness of the solution and the solution itself do not change. Formula (4.186) implies that the following cases may occure concerning the solutions of the inhomogeneous linear system of equations:

Case 1: The system has no solution if any of the numbers $a^{(r-1)}_{r+1}, a^{(r-1)}_{r+2}, \ldots, a^{(r-1)}_m$ differs from zero.

Case 2: The system has a solution if $a^{(r-1)}_{r+1} = a^{(r-1)}_{r+2} = \ldots = a^{(r-1)}_m = 0$ is valid. Then there are two cases:

a) $r = n$: The solution is unique.

b) $r < n$: The solution is not unique; $n - r$ unknowns can be chosen as free parameters.

If the system has a solution, then the unknowns are to be determined in a successive way starting with the last row of the system of equations with the matrix in row echelon form (4.186).

■ **A:** $\begin{aligned} x_1 + 2x_2 + 3x_3 + 4x_4 &= -2 \\ 2x_1 + 3x_2 + 4x_3 + x_4 &= 2 \\ 3x_1 + 4x_2 + x_3 + 2x_4 &= 2 \\ 4x_1 + x_2 + 2x_3 + 3x_4 &= -2 \end{aligned}$ After three Gauss steps the augmented matrix of coefficients has the form $\left(\begin{array}{cccc|c} 1 & 2 & 3 & 4 & -2 \\ 0 & -1 & -2 & -7 & 6 \\ 0 & 0 & -4 & 4 & -4 \\ 0 & 0 & 0 & 40 & -40 \end{array}\right)$.

The solution is unique and from the corresponding system of equations with a triangular matrix follows: $x_4 = -1$, $x_3 = 0$, $x_2 = 1$, $x_1 = 0$.

■ **B:** $\begin{aligned} -x_1 - 3x_2 - 12x_3 &= -5 \\ -x_1 + 2x_2 + 5x_3 &= 2 \\ 5x_2 + 17x_3 &= 7 \\ 3x_1 - x_2 + 2x_3 &= 1 \\ 7x_1 - 4x_2 - x_3 &= 0 \end{aligned}$ After two Gauss steps the augmented matrix of coefficients has the form $\left(\begin{array}{ccc|c} -1 & -3 & -12 & -5 \\ 0 & 5 & 17 & 7 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right)$.

There is a solution but it is not unique. Choosing one unknown as a free parameter, for instance $x_3 = t$ $(-\infty < t < \infty)$, and one gets $x_3 = t$, $x_2 = \frac{7}{5} - \frac{17}{5}t$, $x_1 = \frac{4}{5} - \frac{9}{5}t$.

4.5.3 Overdetermined Linear Systems of Equations

4.5.3.1 Overdetermined Linear Systems of Equations and Linear Least Squares Problems

1. Overdetermined System of Equations

Consider the linear system of equations

$$\mathbf{A}\underline{\mathbf{x}} = \underline{\mathbf{b}} \tag{4.187}$$

with the rectangular matrix of coefficients $\mathbf{A} = (a_{ij})$ $(i = 1, 2, \ldots, m;\ j = 1, 2, \ldots, n;\ m > n)$. The matrix $\mathbf{A}$ and the vector $\underline{\mathbf{b}} = (b_1, b_2, \ldots, b_m)^{\mathrm{T}}$ on the right-hand side are given, and the vector $\underline{\mathbf{x}} = (x_1, x_2, \ldots, x_n)^{\mathrm{T}}$ is unknown. Because $m > n$ holds this system is called an *over-determined system*. One can tell the existence and uniqueness of the solution and sometimes also the solution, for instance by pivoting.

2. Linear Least Squares Problem

If (4.187) is the mathematical model representing a practical problem (i.e., $\mathbf{A}$, $\underline{\mathbf{b}}$, $\underline{\mathbf{x}}$ are reals), then because of measuring or other errors it is impossible to find an exact solution of (4.187) such that it satisfies all of the equations. Substituting any vector $\underline{\mathbf{x}}$ there will be a *residual vector* $\underline{\mathbf{r}} = (r_1, r_2, \ldots, r_m)^{\mathrm{T}}$ given as

$$\underline{\mathbf{r}} = \mathbf{A}\underline{\mathbf{x}} - \underline{\mathbf{b}}, \quad \underline{\mathbf{r}} \neq \underline{\mathbf{0}}. \tag{4.188}$$

In this case $\underline{\mathbf{x}}$ is to be determined to make the norm of the residual vector $\underline{\mathbf{r}}$ as small as possible. Suppose now $\mathbf{A}$, $\underline{\mathbf{b}}$, $\underline{\mathbf{x}}$ are real. If considering the Euclidean norm, then

$$\sum_{i=1}^{m} r_i^2 = \underline{\mathbf{r}}^{\mathrm{T}}\underline{\mathbf{r}} = (\mathbf{A}\underline{\mathbf{x}} - \underline{\mathbf{b}})^{\mathrm{T}}(\mathbf{A}\underline{\mathbf{x}} - \underline{\mathbf{b}}) = \min \tag{4.189}$$

must be valid, i.e., the *residual sum of squares* must be minimal. Gauss already had this idea. The formula (4.189) is called a *linear least squares problem*. The *norm* $||\underline{\mathbf{r}}|| = \sqrt{\underline{\mathbf{r}}^{\mathrm{T}}\underline{\mathbf{r}}}$ of the residual vector $\underline{\mathbf{r}}$ is called the *residue*.

3. Gauss Transformation

The vector $\underline{\mathbf{x}}$ is the solution of (4.189) if the residual vector $\underline{\mathbf{r}}$ is orthogonal to every column of $\mathbf{A}$. That is:

$$\mathbf{A}^{\mathrm{T}}\underline{\mathbf{r}} = \mathbf{A}^{\mathrm{T}}(\mathbf{A}\underline{\mathbf{x}} - \underline{\mathbf{b}}) = \underline{\mathbf{0}} \qquad \text{or} \qquad \mathbf{A}^{\mathrm{T}}\mathbf{A}\underline{\mathbf{x}} = \mathbf{A}^{\mathrm{T}}\underline{\mathbf{b}}. \tag{4.190}$$

Equation (4.190) is actually a linear system of equations with a square matrix of coefficients. One refers to it as the *system of normal equations*. It has dimension n. The transition from (4.187) to (4.190) is called *Gauss transformation*. The matrix $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ is symmetric.

If the matrix $\mathbf{A}$ has the rank n (because $m > n$ all columns of $\mathbf{A}$ are independent), then the matrix $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ is positive definite and also regular, i.e., the system of normal equations has a unique solution if the rank of $\mathbf{A}$ is equal to the number of unknowns.

4.5.3.2 Suggestions for Numerical Solutions of Least Squares Problems

1. Cholesky Method

Because the matrix $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ is symmetric and positive definite in the case rank $(\mathbf{A}) = n$, in order to solve the normal system of equations one can use the Cholesky method (see 19.2.1.2, p. 958). Unfortunately this algorithm is numerically fairly unstable although it works fairly well in the cases of a "big" residue $||\underline{\mathbf{r}}||$ and a "small" solution $||\underline{\mathbf{x}}||$.

2. Householder Method

Numerically useful procedures in order to solve the least squares problem are the *orthogonalization methods* which are based on the decomposition $\mathbf{A} = \mathbf{QR}$. Especially useful is the *Householder method*, where $\mathbf{Q}$ is an orthogonal matrix of size (m, m) and $\mathbf{R}$ is a triangular matrix of size (m, n) (see 4.1.2, **11.**, p. 271).

3. Regularized Problem

In the case of *rank deficiency*, i.e., if rank $(\mathbf{A}) < n$ holds, then the normal system of equations no longer has a unique solution, and the orthogonalization method gives useless results. Then instead of (4.189) the so-called *regularized problem* is considered:

$$\underline{\mathbf{r}}^{\mathrm{T}}\underline{\mathbf{r}} + \alpha\underline{\mathbf{x}}^{\mathrm{T}}\underline{\mathbf{x}} = \min! \qquad (4.191)$$

Here $\alpha > 0$ is a *regularization parameter*. The normal equations for (4.191) are:

$$(\mathbf{A}^{\mathrm{T}}\mathbf{A} + \alpha\mathbf{I})\underline{\mathbf{x}} = \mathbf{A}^{\mathrm{T}}\underline{\mathbf{b}}. \qquad (4.192)$$

The matrix of coefficients of this linear system of equations is positive definite and regular for $\alpha > 0$, but the appropriate choice of the regularization parameter α is a difficult problem (see [4.7]).

4.6 Eigenvalue Problems for Matrices

4.6.1 General Eigenvalue Problem

Let $\mathbf{A}$ and $\mathbf{B}$ be two square matrices of size (n, n). Their elements can be real or complex numbers. The *general eigenvalue problem* is to determine the numbers λ and the corresponding vectors $\underline{\mathbf{x}} \neq \underline{\mathbf{0}}$ satisfying the equation

$$\mathbf{A}\underline{\mathbf{x}} = \lambda\mathbf{B}\underline{\mathbf{x}}. \qquad (4.193)$$

The number λ is called an *eigenvalue*, the vector $\underline{\mathbf{x}}$ an *eigenvector* corresponding to λ. An eigenvector is determined up to a constant factor, because if $\underline{\mathbf{x}}$ is an eigenvector corresponding to λ, so is $c\underline{\mathbf{x}}$ ($c =$ constant) as well. In the special case when $\mathbf{B} = \mathbf{I}$ holds, where $\mathbf{I}$ is the unit matrix of order n, i.e.,

$$\mathbf{A}\underline{\mathbf{x}} = \lambda\underline{\mathbf{x}} \quad \text{or} \quad (\mathbf{A} - \lambda\mathbf{I})\underline{\mathbf{x}} = \underline{\mathbf{0}}, \qquad (4.194)$$

the problem is called the *special eigenvalue problem*. It occures very often in practical problems, especially with a symmetric matrix $\mathbf{A}$, and so it is to be discussed later in detail. More information about the general eigenvalue problem can be found in the literature (see [4.16]).

4.6.2 Special Eigenvalue Problem

4.6.2.1 Characteristic Polynomial

The eigenvalue equation (4.194) yields a homogeneous system of equations which has non-trivial solutions $\underline{\mathbf{x}} \neq \underline{\mathbf{0}}$ only if

$$\det(\mathbf{A} - \lambda\mathbf{I}) = 0. \tag{4.195a}$$

By the expansion of $\det(\mathbf{A} - \lambda\,\mathbf{I})$ one gets

$$\det(\mathbf{A} - \lambda\mathbf{I}) = \begin{vmatrix} a_{11}-\lambda & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22}-\lambda & a_{23} & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nn}-\lambda \end{vmatrix}$$

$$= P_n(\lambda) = (-1)^n\lambda^n + a_{n-1}\lambda^{n-1} + \cdots + a_1\lambda + a_0 = 0. \tag{4.195b}$$

So the determination of the eigenvalues is equivalent to the solution of a polynomial equation. This equation is called the *characteristic equation*; the polynomial $P_n(\lambda)$ is the *characteristic polynomial*. Its roots are the eigenvalues of the matrix $\mathbf{A}$. For an arbitrary square matrix $\mathbf{A}$ of size (n, n) the following statements hold:

Case 1: The matrix $\mathbf{A}_{(n,n)}$ has exactly n eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$, because a polynomial of degree n has n roots if they are considered with their multiplicity. The eigenvalues of a real symmetric matrix are real numbers, in other cases the eigenvalues can also be complex.

Case 2: If all the n eigenvalues are different, then the matrix $\mathbf{A}_{(n,n)}$ has exactly n linearly independent eigenvectors $\underline{\mathbf{x}}_i$ as the solutions of the equation system (4.194) with $\lambda = \lambda_i$.

Case 3: If λ_i has multiplicity n_i among the eigenvalues, and the rank of the matrix $\mathbf{A}_{(n,n)} - \lambda_i\mathbf{I}$ is equal to r_i, then the number of linearly independent eigenvectors corresponding to λ_i is equal to the so-called *nullity* $n - r_i$ of the matrix of coefficients. The inequality $1 \leq n - r_i \leq n_i$ holds, i.e., for a real or complex quadratic matrix $\mathbf{A}_{(n,n)}$ there are at least one and at most n real or complex linearly independent eigenvectors.

■ **A:** $\begin{pmatrix} 2 & -3 & 1 \\ 3 & 1 & 3 \\ -5 & 2 & -4 \end{pmatrix}$, $\det(\mathbf{A} - \lambda\mathbf{I}) = \begin{vmatrix} 2-\lambda & -3 & 1 \\ 3 & 1-\lambda & 3 \\ -5 & 2 & -4-\lambda \end{vmatrix} = -\lambda^3 - \lambda^2 + 2\lambda = 0.$

The eigenvalues are $\lambda_1 = 0, \lambda_2 = 1, \lambda_3 = -2$. The eigenvectors are determined from the corresponding homogeneous linear system of equations.

• $\lambda_1 = 0$: $\begin{aligned} 2x_1 - 3x_2 + x_3 &= 0 \\ 3x_1 + x_2 + 3x_3 &= 0 \\ -5x_1 + 2x_2 - 4x_3 &= 0. \end{aligned}$

One gets for instance by pivoting: x_1 arbitrary, $x_2 = \frac{3}{10}x_1, x_3 = -2x_1 + 3x_2 = -\frac{11}{10}x_1$. Choosing $x_1 = 10$ the eigenvector is $\underline{\mathbf{x_1}} = C_1 \begin{pmatrix} 10 \\ 3 \\ -11 \end{pmatrix}$, where $C_1 \neq 0$ is an arbitrary constant.

• $\lambda_2 = 1$: The corresponding homogeneous system yields: x_3 is arbitrary, $x_2 = 0, x_1 = 3x_2 - x_3 = -x_3$. Choosing $x_3 = 1$ the eigenvector is $\underline{\mathbf{x_2}} = C_2 \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}$, where $C_2 \neq 0$ is an arbitrary constant.

• $\lambda_3 = -2$: The corresponding homogeneous system yields: x_2 is arbitrary, $x_1 = \frac{4}{3}x_2$, $x_3 = -4x_1 + 3x_2 = -\frac{7}{3}x_2$. Choosing $x_2 = 3$ the eigenvector is $\underline{\mathbf{x_3}} = C_3 \begin{pmatrix} 4 \\ 3 \\ -7 \end{pmatrix}$, where $C_3 \neq 0$ is an arbitrary

constant.

■ **B:** $\begin{pmatrix} 3 & 0 & -1 \\ 1 & 4 & 1 \\ -1 & 0 & 3 \end{pmatrix}$, $\det(\mathbf{A} - \lambda\mathbf{I}) = \begin{vmatrix} 3-\lambda & 0 & -1 \\ 1 & 4-\lambda & 1 \\ -1 & 0 & 3-\lambda \end{vmatrix} = -\lambda^3 + 10\lambda^2 - 32\lambda + 32 = 0.$

The eigenvalues are $\lambda_1 = 2$, $\lambda_2 = \lambda_3 = 4$.

• $\lambda_1 = 2$: One obtains x_3 is arbitrary, $x_2 = -x_3, x_1 = x_3$ and chooses for instance $x_3 = 1$. So the corresponding eigenvector is $\underline{\mathbf{x_1}} = C_1 \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}$, where $C_1 \neq 0$ is an arbitrary constant.

• $\lambda_2 = \lambda_3 = 4$: One obtains x_2, x_3 are arbitrary, $x_1 = -x_3$. There are two linearly independent eigenvectors, e.g., for $x_2 = 1, x_3 = 0$ and $x_2 = 0$, $x_3 = 1$: $\underline{\mathbf{x_2}} = C_2 \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$, $\underline{\mathbf{x_3}} = C_3 \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}$, where $C_2 \neq 0$, $C_3 \neq 0$ are arbitrary constants.

4.6.2.2 Real Symmetric Matrices, Similarity Transformations

In the case of the special eigenvalue problem (4.194) for a real symmetric matrix $\mathbf{A}$ the following statements hold:

1. Properties Concerning the Eigenvalue Problem

1. Number of Eigenvalues The matrix $\mathbf{A}$ has exactly n real eigenvalues λ_i $(i = 1, 2, \ldots, n)$, counting them by their multiplicity.

2. Orthogonality of the Eigenvectors The eigenvectors $\underline{\mathbf{x}}_i$ and $\underline{\mathbf{x}}_j$ corresponding to different eigenvalues $\lambda_i \neq \lambda_j$ are orthogonal to each other, i.e., for the scalar product of $\underline{\mathbf{x}}_i$ and $\underline{\mathbf{x}}_j$

$$\underline{\mathbf{x}}_i^{\mathrm{T}}\, \underline{\mathbf{x}}_j = (\underline{\mathbf{x}}_i, \underline{\mathbf{x}}_j) = 0 \tag{4.196}$$

is valid.

3. Matrix with an Eigenvalue of Multiplicity *p* For an eigenvalue which has multiplicity p $(\lambda = \lambda_1 = \lambda_2 = \ldots = \lambda_p)$, there exist p linearly independent eigenvectors $\underline{\mathbf{x}}_1, \underline{\mathbf{x}}_2, \ldots, \underline{\mathbf{x}}_p$. Because of (4.194) all the non-trivial linear combinations of them are also eigenvectors corresponding to λ. Using the Gram–Schmidt orthogonalization process one can choose p of them such that they are orthogonal to each other.
Summarizing: The matrix $\mathbf{A}$ has exactly n real orthogonal eigenvectors.

■ $\mathbf{A} = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}$, $\det(\mathbf{A} - \lambda\,\mathbf{I}) = -\lambda^3 + 3\lambda + 2 = 0$. The eigenvalues are $\lambda_1 = \lambda_2 = -1$ and $\lambda_3 = 2$.

• $\lambda_1 = \lambda_2 = -1$: From the corresponding homogenous system of equations one gets: x_1 is arbitrary, x_2 is arbitrary, $x_3 = -x_1 - x_2$. Choosing first $x_1 = 1$, $x_2 = 0$ then $x_1 = 0$, $x_2 = 1$ one gets the linearly independent eigenvectors $\underline{\mathbf{x}}_1 = C_1 \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}$ and $\underline{\mathbf{x}}_2 = C_2 \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix}$, where $C_1 \neq 0$ and $C_2 \neq 0$ are arbitrary constants.

• $\lambda_3 = 2$: One gets: x_1 is arbitrary, $x_2 = x_1$, $x_3 = x_1$, and choosing for instance $x_1 = 1$ one gets the eigenvector $\underline{\mathbf{x}}_3 = C_3 \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$, where $C_3 \neq 0$ is an arbitrary constant. The matrix $\mathbf{A}$ is symmetric, so the eigenvectors corresponding to different eigenvalues are orthogonal.

4. Gram-Schmidt Orthogonalization Process Let V_n be an arbitrary n-dimensional Euclidean vector space. Let the vectors $\underline{\mathbf{x}}_1, \underline{\mathbf{x}}_2, \ldots, \underline{\mathbf{x}}_n \in V_n$ be linearly independent. Then there exists an or-

thogonal system of vectors $\underline{\mathbf{y}}_1, \underline{\mathbf{y}}_2, \ldots, \underline{\mathbf{y}}_n \in V_n$ which can be obtained by the vectors $\underline{\mathbf{x}}_i$ as follows:

$$\underline{\mathbf{y}}_1 = \underline{\mathbf{x}}_1,\ \underline{\mathbf{y}}_k = \underline{\mathbf{x}}_k - \sum_{i=1}^{k-1} \frac{(\underline{\mathbf{x}}_k, \underline{\mathbf{y}}_i)}{(\underline{\mathbf{y}}_i, \underline{\mathbf{y}}_i)} \underline{\mathbf{y}}_i \qquad (k = 2, 3, \ldots, n). \tag{4.197}$$

Remarks:

1. Here $(\underline{\mathbf{x}}_k,\ \underline{\mathbf{y}}_i) = \underline{\mathbf{x}}_k^{\mathrm{T}} \underline{\mathbf{y}}_i$ is the scalar product of the vectors $\underline{\mathbf{x}}_k$ and $\underline{\mathbf{y}}_i$.

2. Corresponding to the orthogonal system of the vectors $\underline{\mathbf{y}}_1, \underline{\mathbf{y}}_2, \ldots, \underline{\mathbf{y}}_n$ one gets the orthonormal system $\tilde{\underline{\mathbf{x}}}_1, \tilde{\underline{\mathbf{x}}}_2, \ldots, \tilde{\underline{\mathbf{x}}}_n$ with $\tilde{\underline{\mathbf{x}}}_1 = \dfrac{\underline{\mathbf{y}}_1}{||\underline{\mathbf{y}}_1||},\ \tilde{\underline{\mathbf{x}}}_2 = \dfrac{\underline{\mathbf{y}}_2}{||\underline{\mathbf{y}}_2||}, \ldots,\ \tilde{\underline{\mathbf{x}}}_n = \dfrac{\underline{\mathbf{y}}_n}{||\underline{\mathbf{y}}_n||}$, where $||\underline{\mathbf{y}}_i|| = \sqrt{(\underline{\mathbf{y}}_i, \underline{\mathbf{y}}_i)}$ is the Euclidean norm of the vector $\underline{\mathbf{y}}_i$.

■ $\underline{\mathbf{x}}_1 = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix},\ \underline{\mathbf{x}}_2 = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix},\ \underline{\mathbf{x}}_3 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$. From here it follows:

$$\underline{\mathbf{y}}_1 = \underline{\mathbf{x}}_1 = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} \text{ and } \tilde{\underline{\mathbf{x}}}_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix};\ \underline{\mathbf{y}}_2 = \underline{\mathbf{x}}_2 - \frac{(\underline{\mathbf{x}}_2\,,\,\underline{\mathbf{y}}_1)}{(\underline{\mathbf{y}}_1\,,\,\underline{\mathbf{y}}_1)} \underline{\mathbf{y}}_1 = \begin{pmatrix} 1 \\ -1/2 \\ 1/2 \end{pmatrix} \text{ and } \tilde{\underline{\mathbf{x}}}_2 = \frac{1}{\sqrt{6}} \begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix};$$

$$\underline{\mathbf{y}}_3 = \underline{\mathbf{x}}_3 - \frac{(\underline{\mathbf{x}}_3, \underline{\mathbf{y}}_1)}{(\underline{\mathbf{y}}_1, \underline{\mathbf{y}}_1)} \underline{\mathbf{y}}_1 - \frac{(\underline{\mathbf{x}}_3, \underline{\mathbf{y}}_2)}{(\underline{\mathbf{y}}_2, \underline{\mathbf{y}}_2)} \underline{\mathbf{y}}_2 = \begin{pmatrix} 2/3 \\ 2/3 \\ -2/3 \end{pmatrix} \text{ and } \tilde{\underline{\mathbf{x}}}_3 = \frac{1}{\sqrt{3}} \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix}.$$

2. Transformation of Principal Axes, Similarity Transformation

For every real symmetric matrix $\mathbf{A}$, there is an orthogonal matrix $\mathbf{U}$ and a diagonal matrix $\mathbf{D}$ such that

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}^{\mathrm{T}}. \tag{4.198}$$

The diagonal elements of $\mathbf{D}$ are the eigenvalues of $\mathbf{A}$, and the columns of $\mathbf{U}$ are the corresponding normed eigenvectors. From (4.198) it is obvious that

$$\mathbf{D} = \mathbf{U}^{\mathrm{T}}\mathbf{A}\mathbf{U}. \tag{4.199}$$

Transformation (4.199) is called the *transformation of principal axes*. In this way $\mathbf{A}$ is reduced to a diagonal matrix (see also 4.1.2, **2.**, p. 270).

If the square matrix $\mathbf{A}$ (not necessarily symmetric) is transformed by a square regular matrix $\mathbf{G}$ such a way that

$$\mathbf{G}^{-1}\mathbf{A}\,\mathbf{G} = \tilde{\mathbf{A}} \tag{4.200}$$

then it is called a *similarity transformation*. The matrices $\mathbf{A}$ and $\tilde{\mathbf{A}}$ are called *similar* and they have the following properties:

1. The matrices $\mathbf{A}$ and $\tilde{\mathbf{A}}$ have the same eigenvalues, i.e., the similarity transformation does not affect the eigenvalues.

2. If $\mathbf{A}$ is symmetric and $\mathbf{G}$ is orthogonal, then $\tilde{\mathbf{A}}$ is symmetric, too:

$$\tilde{\mathbf{A}} = \mathbf{G}^{\mathrm{T}}\mathbf{A}\,\mathbf{G} \qquad \text{with} \qquad \mathbf{G}^{\mathrm{T}}\mathbf{G} = \mathbf{I}. \tag{4.201}$$

The relation (4.201) is called an *orthogonal-similarity transformation*. In this context (4.199) means that a real symmetric matrix $\mathbf{A}$ can be transformed orthogonally similar to a real diagonal form $\mathbf{D}$.

4.6.2.3 Transformation of Principal Axes of Quadratic Forms

1. Real Quadratic Form, Definition

A real quadratic form Q of variables $x_1, x_2, \ldots, x_n$ has the form

$$Q = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j = \underline{\mathbf{x}}^{\mathrm{T}}\mathbf{A}\underline{\mathbf{x}}, \tag{4.202}$$

where $\underline{\mathbf{x}} = (x_1, x_2, \dots, x_n)^{\mathrm{T}}$ is the vector of real variables and the matrix $\mathbf{A} = (a_{ij})$ is a real symmetric matrix.
The form Q is called *positive definite* or *negative definite*, if it takes only positive or only negative values respectively, and it takes the zero value only in the case $x_1 = x_2 = \dots = x_n = 0$.
The form Q is called *positive* or *negative semidefinite*, if it takes non-zero values only with the sign according to its name, but it can take the zero value for non-zero vectors, too.
A real quadratic form is called *indefinite* if it takes both positive and negative values. According to the behavior of Q the associated real symmetric matrix $\mathbf{A}$ is called positive or negative definite, semidefinite or indefinite.

2. Real Positive Definite Quadratic Form, Properties

1. In a real positive definite quadratic form Q all elements of the main diagonal of the corresponding real symmetric matrix $\mathbf{A}$ are positive, i.e.,

$$a_{ii} > 0 \quad (i = 1, 2, \dots, n) \tag{4.203}$$

holds. (4.203) represents a very important property of positive definite matrices.

2. A real quadratic form Q is positive definite if and only if all eigenvalues of the corresponding matrix $\mathbf{A}$ are positive.

3. Suppose the rank of the matrix $\mathbf{A}$ corresponding to the real quadratic form $Q = \underline{\mathbf{x}}^{\mathrm{T}}\mathbf{A}\underline{\mathbf{x}}$ is equal to r. Then the quadratic form can be transformed by a linear transformation

$$\underline{\mathbf{x}} = \mathbf{C}\underline{\tilde{\mathbf{x}}} \tag{4.204}$$

into a sum of pure quadratic terms, into the so-called *normal form*

$$Q = \underline{\tilde{\mathbf{x}}}^{\mathrm{T}}\mathbf{K}\underline{\tilde{\mathbf{x}}} = \sum_{i=1}^{r} p_i \tilde{x}_i{}^2 \tag{4.205}$$

where $p_i = (\operatorname{sign} \lambda_i) k_i$ and $k_1, k_2, \dots, k_r$ are arbitrary, previously given, positive constants.

Remark: Regardless of the non-singular transformation (4.204) that transforms the real quadratic form of rank r into the normal form (4.205), the number p of positive coefficients and the number $q = r - p$ of negative coefficients among the p_i of the normal form are invariant (the *inertia theorem of Sylvester*). The value p is called the *index of inertia of the quadratic form.*

3. Generation of the Normal Form

A practical method to use the transformation (4.205) follows from the transformation of principal axes (4.199). First it is to perform a rotation on the coordinate system by the orthogonal matrix $\mathbf{U}$, whose columns are the eigenvectors of $\mathbf{A}$ (i.e., the directions of the axes of the new coordinate system are the directions of the eigenvectors). This gives the form

$$Q = \underline{\tilde{\mathbf{x}}}^{\mathrm{T}}\mathbf{L}\underline{\tilde{\mathbf{x}}} = \sum_{i=1}^{r} \lambda_i \tilde{x}_i{}^2. \tag{4.206}$$

Here $\mathbf{L}$ is a diagonal matrix with the eigenvalues of $\mathbf{A}$ in the diagonal. Then a dilatation is performed by the diagonal matrix $\mathbf{D}$ whose diagonal elements are $d_i = \sqrt{\dfrac{k_i}{|\lambda_i|}}$. The whole transformation is now given by the matrix

$$\mathbf{C} = \mathbf{U}\,\mathbf{D}, \tag{4.207}$$

and one gets:

$$\begin{aligned} Q &= \underline{\tilde{\mathbf{x}}}^{\mathrm{T}}\mathbf{A}\underline{\tilde{\mathbf{x}}} = (\mathbf{U}\,\mathbf{D}\underline{\tilde{\mathbf{x}}})^{\mathrm{T}}\mathbf{A}(\mathbf{U}\,\mathbf{D}\underline{\tilde{\mathbf{x}}}) = \underline{\tilde{\mathbf{x}}}^{\mathrm{T}}(\mathbf{D}^{\mathrm{T}}\mathbf{U}^{\mathrm{T}}\mathbf{A}\,\mathbf{U}\,\mathbf{D})\underline{\tilde{\mathbf{x}}} \\ &= \underline{\tilde{\mathbf{x}}}^{\mathrm{T}}\mathbf{D}^{\mathrm{T}}\mathbf{L}\,\mathbf{D}\underline{\tilde{\mathbf{x}}} = \underline{\tilde{\mathbf{x}}}^{\mathrm{T}}\mathbf{K}\underline{\tilde{\mathbf{x}}}. \end{aligned} \tag{4.208}$$

Remark: The transformation of principal axes of quadratic forms plays an essential role at the classification of curves and surfaces of second order (see 3.5.2.11, p. 206 and 3.5.3.14, p. 228).

4. Jordan Normal Form

Let $\mathbf{A}$ be an arbitrary real or complex (n, n) matrix. Then there exists a non-singular matrix $\mathbf{T}$ such that

$$\mathbf{T}^{-1}\mathbf{A}\mathbf{T} = \mathbf{J} \tag{4.209}$$

holds, where $\mathbf{J}$ is called the *Jordan matrix* or *Jordan normal form* of $\mathbf{A}$. The Jordan matrix has a block diagonal structure of the form (4.210), where the elemnts $\mathbf{J}_j$ of $\mathbf{J}$ are called *Jordan blocks*:

$$\mathbf{J} = \begin{pmatrix} \mathbf{J}_1 & & & & \\ & \mathbf{J}_2 & & & \mathbf{O} \\ & & \ddots & & \\ \mathbf{O} & & & \mathbf{J}_{k-1} & \\ & & & & \mathbf{J}_k \end{pmatrix}. \tag{4.210}$$

$$\mathbf{J} = \begin{pmatrix} \lambda_1 & & & & \\ & \lambda_2 & & & \mathbf{O} \\ & & \ddots & & \\ \mathbf{O} & & & \lambda_{n-1} & \\ & & & & \lambda_n \end{pmatrix}. \tag{4.211}$$

They have the following structure:

1. If $\mathbf{A}$ has only single eigenvalues λ_j, then $\mathbf{J}_j = \lambda_j$ and $k = n$, i.e., $\mathbf{J}$ is a diagonal matrix (4.211).

2. If λ_j is an eigenvalue of multiplicity p_j, then there are one or more blocks of the form (4.212) where the sum of the sizes of all such blocks is equal to p_j and $\sum_{j=1}^{k} p_j = n$ holds. The exact structure of a Jordan block depends on the structure of the elementary divisors of the characteristic matrix $\mathbf{A} - \lambda\mathbf{I}$.

$$\mathbf{J}_j = \begin{pmatrix} \lambda_j & 1 & & & \\ & \lambda_j & 1 & & \mathbf{O} \\ & & \ddots & \ddots & \\ & & & \lambda_j & 1 \\ \mathbf{O} & & & & \lambda_j \end{pmatrix}, \tag{4.212}$$

For further information see [4.15], [19.16] vol. 1.

4.6.2.4 Suggestions for the Numerical Calculations of Eigenvalues

1. Eigenvalues can be calculated as the roots of the characteristic equation (4.195b) (see examples on p. 315). In order to get them the coefficients a_i $(i = 0, 1, 2, \ldots, n-1)$ of the characteristic polynomial of the matrix $\mathbf{A}$ must be determined. However, one should avoid this method of calculation, because this procedure is extremely unstable, i.e., small changes in the coefficients a_i of the polynomial result in big changes in the roots λ_j.

2. There are many algorithms for the solution of the eigenvalue problem of symmetric matrices. Two types can be distinguished (see [4.7]):

a) Transformation methods, for instance the Jacobi method, Householder tridiagonalization, QR algorithm.

b) Iterative methods, for instance vector iteration, the Rayleigh-Ritz algorithm, inverse iteration, the Lanczos method, the bisection method. As an example the *power method of Mises* is discussed here.

3. The Power Method of Mises Assume that $\mathbf{A}$ is real and symmetric and has a unique dominant eigenvalue. This iteration method determines this eigenvalue and the associated eigenvector. Let the dominant eigenvalue be denoted by λ_1, that is,

$$|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \cdots \geq |\lambda_n|. \tag{4.213}$$

Let $\underline{\mathbf{x}}_1, \underline{\mathbf{x}}_2, \ldots, \underline{\mathbf{x}}_n$ be the associated linearly independent eigenvectors. Then:

1. $\mathbf{A}\underline{\mathbf{x}}_i = \lambda_i \underline{\mathbf{x}}_i \quad (i = 1, 2, \ldots, n).$ (4.214)

2. Each element $\underline{\mathbf{x}} \in \mathbb{R}^n$ can be expressed as a linear combination of these eigenvectors $\underline{\mathbf{x}}_i$:

$$\underline{\mathbf{x}} = c_1\underline{\mathbf{x}}_1 + c_2\underline{\mathbf{x}}_2 + \cdots + c_n\underline{\mathbf{x}}_n \quad (c_i \text{ const}; \ i = 1, 2, \ldots, n). \tag{4.215}$$

Multiplying both sides of (4.215) by $\mathbf{A}$ k times, then using (4.214) follows

$$\mathbf{A}^k\underline{\mathbf{x}} = c_1\lambda_1^k\underline{\mathbf{x}}_1 + c_2\lambda_2^k\underline{\mathbf{x}}_2 + \cdots + c_n\lambda_n^k\underline{\mathbf{x}}_n = \lambda_1^k\left[c_1\underline{\mathbf{x}}_1 + c_2\left(\frac{\lambda_2}{\lambda_1}\right)^k\underline{\mathbf{x}}_2 + \cdots + c_n\left(\frac{\lambda_n}{\lambda_1}\right)^k\underline{\mathbf{x}}_n\right]. \tag{4.216}$$

From this relation and (4.213) one can see that

$$\frac{\mathbf{A}^k \underline{\mathbf{x}}}{\lambda_1^k c_1} \longrightarrow \underline{\mathbf{x}}_1 \text{ as } k \to \infty\,, \quad \text{that is, } \mathbf{A}^k \underline{\mathbf{x}} \approx c_1 \lambda_1^k \underline{\mathbf{x}}_1. \tag{4.217}$$

This is the basis of the following iteration procedure:
Step 1: Select an arbitrary starting vector $\underline{\mathbf{x}}^{(0)} \in \mathbb{R}^n$.
Step 2: Iterative computation of

$$\underline{\mathbf{x}}^{(k+1)} = \mathbf{A}\underline{\mathbf{x}}^{(k)} \quad (k = 0, 1, 2, \ldots; \underline{\mathbf{x}}^{(0)} \text{ is given}). \tag{4.218}$$

From (4.218) and keeping in mind (4.217) follows:

$$\underline{\mathbf{x}}^{(k)} = \mathbf{A}^k \underline{\mathbf{x}}^{(0)} \approx c_1 \lambda_1^k \underline{\mathbf{x}}_1. \tag{4.219}$$

Step 3: From (4.218) and (4.219) it follows that

$$\begin{aligned} \underline{\mathbf{x}}^{(k+1)} &= \mathbf{A}\underline{\mathbf{x}}^{(k)} = \mathbf{A}(\mathbf{A}^k \underline{\mathbf{x}}^{(0)}), \\ \mathbf{A}(\mathbf{A}^k \underline{\mathbf{x}}^{(0)}) &\approx \mathbf{A}(c_1 \lambda_1^k \underline{\mathbf{x}}_1) = c_1 \lambda_1^k (\mathbf{A}\underline{\mathbf{x}}_1), \\ c_1(\lambda_1^k \mathbf{A}\underline{\mathbf{x}}_1) &= \lambda_1 (c_1 \lambda_1^k \underline{\mathbf{x}}_1) \approx \lambda_1 \underline{\mathbf{x}}^{(k)}, \text{ therefore} \\ \underline{\mathbf{x}}^{(k+1)} &\approx \lambda_1 \underline{\mathbf{x}}^{(k)}, \end{aligned} \tag{4.220}$$

that is, for large values of k the consecutive vectors $\underline{\mathbf{x}}^{(k+1)}$ and $\underline{\mathbf{x}}^{(k)}$ differ approximately by a factor λ_1.
Step 4: Relations (4.219) and (4.220) imply for $\underline{\mathbf{x}}_1$ and λ_1:

$$\underline{\mathbf{x}}_1 \approx \underline{\mathbf{x}}^{(k+1)}, \quad \lambda_1 \approx \frac{\left(\underline{\mathbf{x}}^{(k)}, \underline{\mathbf{x}}^{(k+1)}\right)}{\left(\underline{\mathbf{x}}^{(k)}, \underline{\mathbf{x}}^{(k)}\right)}. \tag{4.221}$$

■ For example, let

$$\mathbf{A} = \begin{pmatrix} 2.23 & -1.15 & 1.77 \\ -1.15 & 9.25 & -2.13 \\ 1.77 & -2.13 & 1.56 \end{pmatrix}, \quad \underline{\mathbf{x}}^{(0)} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}.$$

$\underline{\mathbf{x}}^{(0)}$	$\underline{\mathbf{x}}^{(1)}$	$\underline{\mathbf{x}}^{(2)}$	$\underline{\mathbf{x}}^{(3)}$	normalization	$\underline{\mathbf{x}}^{(4)}$	$\underline{\mathbf{x}}^{(5)}$	normalization
1	3.23	14.89	88.27	1	7.58	67.75	1
0	−1.15	−18.12	−208.03	−2.36	−24.93	−256.85	−3.79
0	1.77	10.93	82.00	0.93	8.24	79.37	1.17
λ_1				9.964			10.177

$\underline{\mathbf{x}}^{(6)}$	$\underline{\mathbf{x}}^{(7)}$	normalization	$\underline{\mathbf{x}}^{(8)}$	$\underline{\mathbf{x}}^{(9)}$	normalization
9.66	96.40	1	10.09	102.33	1
− 38.78	−394.09	−4.09	−41.58	−422.49	−4.129
11.67	117.78	1.22	12.38	125.73	1.229
		10.16			$10.161 \approx \lambda_1$

$$\begin{pmatrix} 1 \\ -4.129 \\ 1.229 \end{pmatrix} \approx \underline{\mathbf{x}}_1$$

Remarks:

1. Since eigenvectors are unique only up to a constant multiplier, it is preferable to normalize the vectors $\underline{x}^{(k)}$ as shown in the example.

2. The eigenvalue with the smallest absolute value and the associated eigenvector can be obtained by using the power method of Mises for $\mathbf{A}^{-1}$. If $\mathbf{A}^{-1}$ does not exist, then 0 is this eigenvalue and any vector from the null-space of $\mathbf{A}$ can be selected as an associated eigenvector.

3. The other eigenvalues and the associated eigenvectors of $\mathbf{A}$ can be obtained by repeated application

of the following idea. Select a starting vector which is orthogonal to the known vector $\underline{\mathbf{x}}_1$, and in this subspace λ_2 becomes the dominant eigenvalue that can be obtained by using the power method. In order to obtain λ_3, the starting vector has to be orthogonal to both $\underline{\mathbf{x}}_1$ and $\underline{\mathbf{x}}_2$, and so on. This procedure is known as *matrix deflation*.

4. Based on (4.218) the power method is sometimes called *vector iteration*.

4.6.3 Singular Value Decomposition

1. Singular Values and Singular Vectors Let $\mathbf{A}$ be a real matrix of size (m, n) and its rank be equal to r. The matrices $\mathbf{A}\mathbf{A}^{\mathrm{T}}$ and $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ have r non-zero eigenvalues λ_ν, and they are the same for both of the matrices. The positive square roots $d_\nu = \sqrt{\lambda_\nu}$ $\quad(\nu = 1, 2, \dots, r)$ of the eigenvalues λ_ν of the matrix $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ are called the *singular values* of the matrix $\mathbf{A}$. The corresponding eigenvectors $\underline{\mathbf{u}}_\nu$ of $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ are called *right-singular vectors* of $\mathbf{A}$, the corresponding eigenvectors $\underline{\mathbf{v}}_\nu$ of $\mathbf{A}\mathbf{A}^{\mathrm{T}}$ *left-singular vectors*:

$$\mathbf{A}^{\mathrm{T}}\mathbf{A}\underline{\mathbf{u}}_\nu = \lambda_\nu \underline{\mathbf{u}}_\nu, \quad \mathbf{A}\mathbf{A}^{\mathrm{T}}\underline{\mathbf{v}}_\nu = \lambda_\nu \underline{\mathbf{v}}_\nu \quad (\nu = 1, 2, \dots, r). \tag{4.222a}$$

The relations between the right and left-singular vectors are:

$$\mathbf{A}\underline{\mathbf{u}}_\nu = d_\nu \underline{\mathbf{v}}_\nu, \quad \mathbf{A}^{\mathrm{T}}\underline{\mathbf{v}}_\nu = d_\nu \underline{\mathbf{u}}_\nu. \tag{4.222b}$$

A matrix $\mathbf{A}$ of size (m, n) with rank r has r positive-singular values d_ν $\quad(\nu = 1, 2, \dots, r)$. There exist r orthonormalized right-singular vectors $\underline{\mathbf{u}}_\nu$ and r orthonormalized left-singular vectors $\underline{\mathbf{v}}_\nu$. Furthermore, there exist to the zero-singular value $n - r$ orthonormalized right-singular vectors $\underline{\mathbf{u}}_\nu (\nu = r + 1, \dots, n)$ and $m - r$ orthonormalized left-singular vectors $\underline{\mathbf{v}}_\nu$ $\quad(\nu = r + 1, \dots, m)$. Consequently, a matrix of size (m, n) has n right-singular vectors and m left-singular vectors, and two orthogonal matrices can be made from them (see 4.1.4, **9.**, p. 275):

$$\mathbf{U} = (\underline{\mathbf{u_1}}, \underline{\mathbf{u_2}}, \dots, \underline{\mathbf{u_n}}), \quad \mathbf{V} = (\underline{\mathbf{v_1}}, \underline{\mathbf{v_2}}, \dots, \underline{\mathbf{v_m}}). \tag{4.223}$$

2. Singular-Value Decomposition The representation

$$\mathbf{A} = \mathbf{V}\hat{\mathbf{A}}\mathbf{U}^{\mathrm{T}} \quad (4.224\mathrm{a}) \qquad \text{with } \hat{\mathbf{A}} = \underbrace{\left(\begin{array}{ccccc|ccc} d_1 & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & d_2 & & & 0 & 0 & & \vdots \\ \vdots & & \ddots & & \vdots & \vdots & & \\ & & & & 0 & & & \\ 0 & & \cdots & 0 & d_r & 0 & \cdots & 0 \\ \hline 0 & 0 & & \cdots & 0 & 0 & \cdots & 0 \\ 0 & & & & 0 & \vdots & & \\ \vdots & & & & \vdots & & & \\ 0 & 0 & & \cdots & 0 & 0 & \cdots & 0 \end{array}\right)}_{r \text{ columns} \quad n-r \text{ columns}} \begin{array}{l} \left.\vphantom{\begin{array}{c}1\\1\\1\\1\\1\end{array}}\right\} r \text{ rows} \\ \left.\vphantom{\begin{array}{c}1\\1\\1\\1\end{array}}\right\} m-r \text{ rows} \end{array} \tag{4.224b}$$

is called the *singular-value decomposition* of the matrix $\mathbf{A}$. The matrix $\hat{\mathbf{A}}$, as the matrix $\mathbf{A}$, is of size (m, n) and has only zero elements except the first r diagonal elements $a_{\nu\nu} = d_\nu$ $\quad(\nu = 1, 2, \dots, r)$. The values d_ν are the singular values of $\mathbf{A}$.

Remark: If substituting A^{H} instead of A^{T} and consider unitary matrices U and V instead of orthogonals, then all the statements about singular-value decomposition are valid also for matrices with complex elements.

3. Application Singular-value decomposition can be used to determine the rank of the matrix $\mathbf{A}$ of size (m, n) and to calculate an approximate solution of the over-determined system of equations

$\mathbf{A}\underline{\mathbf{x}} = \underline{\mathbf{b}}$ (see 4.5.3.1, p. 313) with the so-called *regularization method*, i.e., to solve the problem

$$||\mathbf{A}\underline{\mathbf{x}} - \underline{\mathbf{b}}||^2 + \alpha||\underline{\mathbf{x}}||^2 = \sum_{i=1}^{m}\left[\sum_{k=1}^{n} a_{ik}x_k - b_i\right]^2 + \alpha\sum_{k=1}^{n} x_k^2 = \min!\,, \tag{4.225}$$

where $\alpha > 0$ is a regularization parameter.

5 Algebra and Discrete Mathematics

5.1 Logic

5.1.1 Propositional Calculus

1. Propositions

A *proposition* is the mental reflection of a fact, expressed as a sentence in a natural or artificial language. Every proposition is considered to be true or false. This is the *principle of two-valuedness* (in contrast to many-valued or fuzzy logic, see 5.9.1, p. 413). "True" and "false" are called the *truth value* of the proposition and they are denoted by T (or 1) and F (or 0), respectively. The truth values can be considered as *propositional constants*.

2. Propositional Connectives

Propositional logic investigates the truth of *compositions of propositions* depending on the truth of the components. Only the *extensions* of the sentences corresponding to propositions are considered. Thus the truth of a composition depends *only* on that of the components and on the operations applied. So in particular, the truth of the result of the propositional operations

"NOT A" $(\neg A)$, (5.1) "A AND B" $(A \wedge B)$, (5.2)

"A OR B" $(A \vee B)$, (5.3) "IF A, THEN B" $(A \Rightarrow B)$ (5.4)

and

"A IF AND ONLY IF B" $(A \Leftrightarrow B)$ (5.5)

are determined by the truth of the components. Here "logical OR" always means "inclusive OR", i.e., "AND/OR". In the case of implication, for $A \Rightarrow B$ also the following verbal forms are in use:

A implies B, B is necessary for A, A is sufficient for B.

3. Truth Tables

In propositional calculus, the propositions A and B are considered as variables (*propositional variables*) which can have only the values F and T. Then the *truth tables* in **Table 5.1** contain the *truth functions* defining the propositional operations.

Table 5.1 Truth tables of propositional calculus

Negation

A	$\neg A$
F	T
T	F

Conjunction

A	B	$A \wedge B$
F	F	F
F	T	F
T	F	F
T	T	T

Disjunction

A	B	$A \vee B$
F	F	F
F	T	T
T	F	T
T	T	T

Implication

A	B	$A \Rightarrow B$
F	F	T
F	T	T
T	F	F
T	T	T

Equivalence

A	B	$A \Leftrightarrow B$
F	F	T
F	T	F
T	F	F
T	T	T

4. Formulas in Propositional Calculus

Compound expressions (formulas) of propositional calculus can be composed from the propositional variables in terms of a unary operation (negation) and binary operations (conjunction, disjunction, implication and equivalence). These expressions, i.e., the formulas, are defined in an inductive way:

1. Propositional variables and the constants T, F are formulas. (5.6)

2. If A and B are formulas, then $(\neg A)$, $(A \wedge B)$, $(A \vee B)$, $(A \Rightarrow B)$, $(A \Leftrightarrow B)$ (5.7)

I.N. Bronshtein et al., *Handbook of Mathematics*,
DOI 10.1007/978-3-662-46221-8_5

are also formulas.
To simplify formulas parentheses are omitted after introducing *precedence rules*. In the following sequence every propositional operation binds more strongly than the next one in the sequence:

$$\neg, \ \wedge, \ \vee, \ \Rightarrow, \ \Leftrightarrow .$$

Often the notation $\overline{A}$ instead of "$\neg A$" is used, and the symbol $\wedge$ is omitted. By these simplifications, for instance the formula $((A \vee (\neg B)) \Rightarrow ((A \wedge B) \vee C))$ can be rewritten more briefly in the form:

$$A \vee \overline{B} \Rightarrow AB \vee C.$$

5. Truth Functions

Assigning a truth value to every propositional variable of a formula, the assignment is called an *interpretation* of the propositional variables. Using the definitions (truth tables) of propositional operations a truth value can be assigned to a formula for every possible interpretation of the variables. Thus for instance the formula given above determines a truth function of three variables (a *Boolean function* see 5.7.5, p. 413).

A	B	C	$A \vee \overline{B}$	$AB \vee C$	$A \vee \overline{B} \Rightarrow AB \vee C$
F	F	F	T	F	F
F	F	T	T	T	T
F	T	F	F	F	T
F	T	T	F	T	T
T	F	F	T	F	F
T	F	T	T	T	T
T	T	F	T	T	T
T	T	T	T	T	T

■ In this way, every formula with n propositional variables determines an n place (or n ary) truth function, i.e., a function which assigns a truth value to every n tuple of truth values. There are 2^{2^n} n ary truth functions, in particular these are 16 binary ones.

6. Elementary Laws in Propositional Calculus

Two propositional formulas A and B are called *logically equivalent* or *semantically equivalent*, denoted by $A = B$, if they determine the same truth function. Consequently, the logical equivalence of propositional formulas can be checked in terms of truth tables. So there is , e.g., $A \vee \overline{B} \Rightarrow AB \vee C = B \vee C$, i.e., the formula $A \vee \overline{B} \Rightarrow AB \vee C$ does not in fact depend on A, as follows from its truth table above. In particular, there are the following *elementary laws of propositional calculus*:

1. Associative Laws

$$(A \wedge B) \wedge C = A \wedge (B \wedge C), \quad (5.8a) \qquad (A \vee B) \vee C = A \vee (B \vee C). \quad (5.8b)$$

2. Commutative Laws

$$A \wedge B = B \wedge A, \quad (5.9a) \qquad A \vee B = B \vee A. \quad (5.9b)$$

3. Distributive Laws

$$(A \vee B)C = AC \vee BC, \quad (5.10a) \qquad AB \vee C = (A \vee C)(B \vee C). \quad (5.10b)$$

4. Absorption Laws

$$A(A \vee B) = A, \quad (5.11a) \qquad A \vee AB = A. \quad (5.11b)$$

5. Idempotence Laws

$$AA = A, \quad (5.12a) \qquad A \vee A = A. \quad (5.12b)$$

6. Excluded Middle

$$A\overline{A} = \mathrm{F}, \quad (5.13a) \qquad A \vee \overline{A} = \mathrm{T}. \quad (5.13b)$$

7. De Morgan Rules

$$\overline{AB} = \overline{A} \vee \overline{B}, \quad (5.14a) \qquad \overline{A \vee B} = \overline{A}\,\overline{B}. \quad (5.14b)$$

8. Laws for T and F

$$A\mathrm{T} = A, \quad (5.15a) \qquad A \vee \mathrm{F} = A, \quad (5.15b)$$

$$A\mathrm{F} = \mathrm{F}, \quad (5.15c) \qquad A \vee \mathrm{T} = \mathrm{T}, \quad (5.15d)$$

$$\overline{\mathrm{T}} = \mathrm{F}, \quad (5.15e) \qquad \overline{\mathrm{F}} = \mathrm{T}. \quad (5.15f)$$

9. Double Negation

$$\overline{\overline{A}} = A. \quad (5.16)$$

Using the truth tables for implication and equivalence, gives the identities

$$A \Rightarrow B = \overline{A} \vee B \quad (5.17a) \qquad \text{and } A \Leftrightarrow B = AB \vee \overline{A}\,\overline{B}. \quad (5.17b)$$

Therefore implication and equivalence can be expressed in terms of other propositional operations. Laws (5.17a), (5.17b) are applied to reformulate propositional formulas.

■ The identity $A \vee \overline{B} \Rightarrow AB \vee C = B \vee C$ can be verified in the following way: $A \vee \overline{B} \Rightarrow AB \vee C = \overline{A \vee \overline{B}} \vee AB \vee C = \overline{A}\,\overline{\overline{B}} \vee AB \vee C = \overline{A}B \vee AB \vee C = (\overline{A} \vee A)B \vee C = \mathrm{T}B \vee C = B \vee C$.

10. Further Transformations

$$A(\overline{A} \vee B) = AB, \quad (5.18a) \qquad A \vee \overline{A}B = A \vee B, \quad (5.18b)$$

$$(A \vee C)(B \vee \overline{C})(A \vee B) = (A \vee C)(B \vee \overline{C}), \quad (5.18c) \qquad AC \vee B\overline{C} \vee AB = AC \vee B\overline{C}. \quad (5.18d)$$

11. NAND Function and NOR Function As it is known, every propositional formula determines a truth function. Checking the following converse of this statement: Every truth function can be represented as a truth table of a suitable formula in propositional logic. Because of (5.17a) and (5.17b) implication and equivalence can be eliminated from formulas (see also 5.7, p. 395). This fact and the De Morgan rules (5.14a) and (5.14b) imply that one can express every formula, therefore every truth function, in terms of negation and disjunction only, or in terms of negation and conjunction. There are two further binary truth functions of two variables which are suitable to express all the truth functions. They are called the NAND function or Sheffer function (notation "$|$") and the NOR function or Peirce function (notation "$\downarrow$"), with the truth tables given in **Tables 5.2** and **5.3**. Comparison of the truth tables for these operations with the truth tables of conjunction and disjunction makes the terminologies NAND function (NOT AND) and NOR function (NOT OR) clear.

Table 5.2 NAND function

A	B	$A \vert B$
F	F	T
F	T	T
T	F	T
T	T	F

Table 5.3 NOR function

A	B	$A \downarrow B$
F	F	T
F	T	F
T	F	F
T	T	F

7. Tautologies, Inferences in Mathematics

A formula in propositional calculus is called a *tautology* if the value of its truth function is identically the value T. Consequently, two formulas A and B are called logically equivalent if the formula $A \Leftrightarrow B$ is a tautology. Laws of propositional calculus often reflect inference methods used in mathematics. As an example, consider the *law of contraposition*, i.e., the tautology

$$A \Rightarrow B \Leftrightarrow \overline{B} \Rightarrow \overline{A}. \quad (5.19a)$$

This law, which also has the form

$$A \Rightarrow B = \overline{B} \Rightarrow \overline{A}, \quad (5.19b)$$

can be interpreted in this way: To show that B is a consequence of A is the same as showing that $\overline{A}$ is a consequence of $\overline{B}$. The *Indirect proof* (see also 1.1.2.2, p. 5) is based on the following principle: To show

that B is a consequence of A, one supposes B to be false, and under the assumption that A is true, one derives a contradiction. This principle can be formalized in propositional calculus in several ways:

$$A \Rightarrow B = A\overline{B} \Rightarrow \overline{A} \qquad (5.20a) \qquad \text{or} \quad A \Rightarrow B = A\overline{B} \Rightarrow B \quad \text{or} \qquad (5.20b)$$

$$A \Rightarrow B = A\overline{B} \Rightarrow \mathrm{F}. \qquad (5.20c)$$

5.1.2 Formulas in Predicate Calculus

For developing the logical foundations of mathematics one needs a logic which has a stronger expressive power than propositional calculus. To describe the properties of most of the objects in mathematics and the relations between these objects the predicate calculus is needed.

1. Predicates

The objects to be investigated are included into a set, i.e., into the *domain X of individuals (or universe)*, e.g., this domain could be the set $\mathbb{N}$ of the natural numbers. The properties of the individuals, as, e.g., " n is a prime ", and the relations between individuals, e.g., " m is smaller than n ", are considered as *predicates*. An *n place predicate* over the domain X of individual is an assignment P: $X^n \rightarrow \{\mathrm{F,W}\}$, which assigns a truth value to every n tuple of the individuals. So the predicates introduced above on natural numbers are a one-place (or unary) predicate and a two-place (or binary) predicate.

2. Quantifiers

A characteristic feature of predicate logic is the use of *quantifiers*, i.e., that of a *universal quantifier* or *"for every" quantifier* $\forall$ and *existential quantifier* or *"for some" quantifier* $\exists$. If P is a unary predicate, then the sentence "$P(x)$ is true for every x in X" is denoted by $\forall x\, P(x)$ and the sentence" There exists an x in X for which $P(x)$ is true "is denoted by $\exists x\, P(x)$. Applying a quantifier to the unary predicate P, gives a sentence. If for instance $\mathbb{N}$ is the domain of individual of the natural numbers and P denotes the (unary) predicate "n is a prime", then $\forall n\, P(n)$ is a false sentence and $\exists n\, P(n)$ is a true sentence.

3. Formulas in Predicate Calculus

The *formulas in predicate calculus* are defined in an inductive way:

1. If $x_1, \ldots, x_n$ are individual variables (variables running over the domain of individual variables) and P is an n-place predicate symbol, then

$$P(x_1, \ldots, x_n) \ \text{ is a formula } (\textit{elementary formula}). \qquad (5.21)$$

2. If A and B are formulas, then

$$(\neg A),\ (A \wedge B),\ (A \vee B),\ (A \Rightarrow B), (A \Leftrightarrow B),\ (\forall x\ A) \ \text{ and } \ (\exists x\ A) \qquad (5.22)$$

are also formulas.

Considering a propositional variable to be a null-place predicate, the propositional calculus can be considered as a part of predicate calculus. An occurrence of an individual variable x is *bound* in a formula if x is a variable in $\forall x$ or in $\exists x$ or the occurrence of x is in the scope of these types of quantifiers; otherwise an occurrence of x is *free* in this formula. A formula of predicate logic which does not contain any free occurrences of individual variables is called a *closed formula*.

4. Interpretation of Predicate Calculus Formulas

An *interpretation* of predicate calculus is a pair of

- a set (domain of individuals) and
- an assignment, which assigns an n-place predicate to every n-ary predicate symbol.

For every prefixed value of free variables the concept of the truth evaluation of a formula is similar to the propositional case. The truth value of a closed formula is T or F. In the case of a formula containing free variables, one can associate the values of individuals for which the truth evaluation of the formula is true; these values constitute a relation (see 5.2.3, **1.**, p. 331) on the universe (domain of individuals).

■ Let P denote the two-place relation $\leq$ on the domain $\mathbb{N}$ of individuals, where $\mathbb{N}$ is the set of the natural numbers then

• $P(x,y)$ characterizes the set of all the pairs (x,y) of natural numbers with $x \le y$ (two-place or binary relation on $\mathbb{N}$); here x, y are free variables;
• $\forall y\, P(x,y)$ characterizes the subset of $\mathbb{N}$ (unary relation) consisting of the element 0 only; here x is a free variable, y is a bound variable;
• $\exists x\, \forall y\, P(x,y)$ corresponds to the sentence " There is a smallest natural number "; the truth value is true; here x and y are bound variables.

5. Logically Valid Formulas

A formula is called *logically valid* (or a *tautology*) if it is true for every interpretation. The negation of formulas is characterized by the identities below:

$$\neg\forall x\, P(x) = \exists x\, \neg P(x) \quad \text{or} \quad \neg\exists x\, P(x) = \forall x\, \neg P(x). \tag{5.23}$$

Using (5.23) the quantifiers $\forall$ and $\exists$ can be expressed in terms of each other:

$$\forall x\, P(x) = \neg\exists x\, \neg P(x) \quad \text{or} \quad \exists x\, P(x) = \neg\forall x\, \neg P(x). \tag{5.24}$$

Further identities of the predicate calculus are:

$$\forall x\, \forall y\, P(x,y) = \forall y\, \forall x\, P(x,y), \tag{5.25}$$

$$\exists x\, \exists y\, P(x,y) = \exists y\, \exists x\, P(x,y), \tag{5.26}$$

$$\forall x\, P(x) \wedge \forall x\, Q(x) = \forall x\, (P(x) \wedge Q(x)), \tag{5.27}$$

$$\exists x\, P(x) \vee \exists x\, Q(x) = \exists x\, (P(x) \vee Q(x)). \tag{5.28}$$

The following implications are also valid:

$$\forall x\, P(x) \vee \forall x\, Q(x) \Rightarrow \forall x\, (P(x) \vee Q(x)), \tag{5.29}$$

$$\exists x\, (P(x) \wedge Q(x)) \Rightarrow \exists x\, P(x) \wedge \exists x\, Q(x), \tag{5.30}$$

$$\forall x\, (P(x) \Rightarrow Q(x)) \Rightarrow (\forall x\, P(x) \Rightarrow \forall x\, Q(x)), \tag{5.31}$$

$$\forall x\, (P(x) \Leftrightarrow Q(x)) \Rightarrow (\forall x\, P(x) \Leftrightarrow \forall x\, Q(x)), \tag{5.32}$$

$$\exists x\, \forall y\, P(x,y) \Rightarrow \forall y\, \exists x\, P(x,y). \tag{5.33}$$

The converses of these implications are not valid, in particular, one has to be careful with the fact that the quantifiers $\forall$ and $\exists$ do not commute (the converse of the last implication is false).

6. Restricted Quantification

Often it is useful to restrict quantification to a subset of a given set. So, there is considered

$$\forall x \in X\; P(x) \quad \text{as a short notation of} \quad \forall x\, (x \in X \Rightarrow P(x)) \quad \text{and} \tag{5.34}$$

$$\exists x \in X\; P(x) \quad \text{as a short notation of} \quad \exists x\, (x \in X \wedge P(x)). \tag{5.35}$$

5.2 Set Theory

5.2.1 Concept of Set, Special Sets

The founder of set theory is Georg Cantor (1845–1918). The importance of the notion introduced by him became well known only later. Set theory has a decisive role in all branches of mathematics, and today it is an essential tool of mathematics and its applications.

1. Membership Relation

1. Sets and their Elements The fundamental notion of set theory is the membership relation. A *set* A is a collection of certain different things a (objects, ideas, etc.) that belong together for certain reasons. These objects are called the *elements* of the set. One writes "$a \in A$" or "$a \notin A$" to denote "a is an element of A" or "a is not an element of A", respectively. Sets can be given by enumerating their elements in braces, e.g., $M = \{a, b, c\}$ or $U = \{1, 3, 5, \ldots\}$, or by a defining property possessed exactly by the elements of the set. For instance the set U of the odd natural numbers is defined and denoted by $U = \{x \mid x \text{ is an odd natural number}\}$. For number domains the following notation is generally used:

$\mathbb{N} = \{0, 1, 2, \ldots\}$	set of the natural numbers,
$\mathbb{Z} = \{0, 1, -1, 2, -2, \ldots\}$	set of the integers,
$\mathbb{Q} = \left\{ \dfrac{p}{q} \,\middle|\, p, q \in \mathbb{Z} \wedge q \neq 0 \right\}$	set of the rational numbers,
$\mathbb{R}$	set of the real numbers,
$\mathbb{C}$	set of the complex numbers.

2. Principle of Extensionality for Sets Two sets A and B are identical if and only if they have exactly the same elements, i.e.,

$$A = B \Leftrightarrow \forall\, x\,(x \in A \Leftrightarrow x \in B). \tag{5.36}$$

■ The sets $\{3, 1, 3, 7, 2\}$ and $\{1, 2, 3, 7\}$ are the same.
A set contains every element only “once”, even if it is enumerated several times.

2. Subsets

1. Subset If A and B are sets and

$$\forall\, x\,(x \in A \Rightarrow x \in B) \tag{5.37}$$

holds, then A is called a *subset* of B, and this is denoted by $A \subseteq B$. In other words: A is a subset of B, if all elements of A also belong to B. If for $A \subseteq B$ there are some further elements in B such that they are not in A, then A is called a *proper subset* of B, and it is denoted by $A \subset B$ **(Fig. 5.1)**. Obviously, every set is a subset of itself $A \subseteq A$.

■ Suppose $A = \{2, 4, 6, 8, 10\}$ is a set of even numbers and $B = \{1, 2, 3, \ldots, 10\}$ is a set of natural numbers. Since the set A does not contain odd numbers, A is a proper subset of B.

2. Empty Set or Void Set It is important and useful to introduce the notion of *empty set* or *void set*, $\emptyset$, which has no element. Because of the principle of extensionality, there exists only one empty set.

■ **A:** The set $\{x | x \in \mathbb{R} \wedge x^2 + 2x + 2 = 0\}$ is empty.

■ **B:** $\emptyset \subseteq M$ for every set M, i.e., the empty set is a subset of every set M.
For a set A the empty set and A itself are called the *trivial subsets of* A.

3. Equality of Sets Two sets are equal if and only if both are subsets of each other:

$$A = B \Leftrightarrow A \subseteq B \wedge B \subseteq A. \tag{5.38}$$

This fact is very often used to prove that two sets are identical.

4. Power Set The set of all subsets A of a set M is called the *power set* of M and it is denoted by $\mathbb{P}(M)$, i.e., $\mathbb{P}(M) = \{A \mid A \subseteq M\}$.

■ For the set $M = \{a, b, c\}$ the power set is

$$\mathbb{P}(M) = \{\emptyset, \{a\}, \{b\}, \{c\}, \{a, b\}, \{a, c\}, \{b, c\}, \{a, b, c\}\}.$$

It is true that:

a) If a set M has m elements, its power set $\mathbb{P}(M)$ has 2^m elements.

b) For every set M there are $M, \emptyset \in \mathbb{P}(M)$, i.e., M itself and the empty set are elements of the power set of M.

5. Cardinal number The number of elements of a finite set M is called the *cardinal number* of M and it is denoted by $\operatorname{card} M$ or sometimes by $|M|$.
For the the cardinal number of sets with infinitely many elements see 5.2.5, p. 335.

5.2.2 Operations with Sets

1. Venn diagram

The graphical representations of sets and set operations are the so-called *Venn diagrams*, when representing sets by plane figures. So, **Fig. 5.1**, represents the subset relation $A \subseteq B$.

2. Union, Intersection, Complement

By *set operations* new sets can be formed from the given sets in different ways:

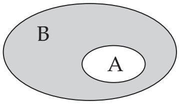


Figure 5.1

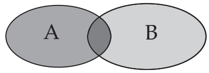


Figure 5.2

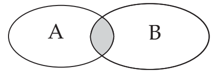


Figure 5.3

1. Union Let A and B be two sets. The *union set* or the *union* (denoted by $A \cup B$) is defined by

$$A \cup B = \{x \mid x \in A \vee x \in B\}, \tag{5.39}$$

in words "A union B" or "A cup B". If A and B are given by the properties E_1 and E_2 respectively, the union set $A \cup B$ has the elements possessing at least one of these properties, i.e., the elements belonging to at least one of the sets. In **Fig. 5.2** the union set is represented by the shaded region.

■ $\{1, 2, 3\} \cup \{2, 3, 5, 6\} = \{1, 2, 3, 5, 6\}$.

2. Intersection Let A and B be two sets. The *intersection set*, *intersection*, *cut* or *cut set* (denoted by $A \cap B$) is defined by

$$A \cap B = \{x \mid x \in A \wedge x \in B\}, \tag{5.40}$$

in words "A intersected by B" or "A cap B". If A and B are given by the properties E_1 and E_2 respectively, the intersection $A \cap B$ has the elements possessing both properties E_1 and E_2, i.e., the elements belonging to both sets. In **Fig. 5.3** the intersection is represented by the shaded region.

■ With the intersection of the sets of divisors $T(a)$ and $T(b)$ of two numbers a and b one can define the greatest common divisor (see 5.4.1.4, p. 373). For $a = 12$ and $b = 18$ holds $T(a) = \{1, 2, 3, 4, 6, 12\}$ and $T(b) = \{1, 2, 3, 6, 9, 18\}$, so $T(12) \cap T(18)$ contains the common divisors, and the greatest common divisor is g.c.d. $(12, 18) = 6$.

3. Disjoint Sets Two sets A and B are called *disjoint* if they have no common element; for them

$$A \cap B = \emptyset \tag{5.41}$$

holds, i.e., their intersection is the empty set.

■ The set of odd numbers and the set of even numbers are disjoint; their intersection is the empty set, i.e.,

$$\{\text{odd numbers}\} \cap \{\text{even numbers}\} = \emptyset.$$

4. Complement Considering only the subsets of a given set M, then the *complementary set* or the *complement* $C_M(A)$ of A with respect to M contains all the elements of M not belonging to A:

$$C_M(A) = \{x \mid x \in M \wedge x \notin A\}, \tag{5.42}$$

in words "complement of A with respect to M", and M is called the *fundamental set* or sometimes the *universal set*. If the fundamental set M is obvious from the considered problem, then the notation $\overline{A}$ is also used for the complementary set. In **Fig. 5.4** the complement $\overline{A}$ is represented by the shaded region.

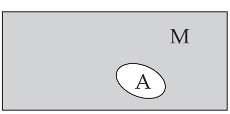


Figure 5.4

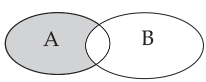


Figure 5.5

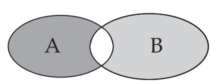


Figure 5.6

3. Fundamental Laws of Set Algebra

These set operations have analoguous properties to the operations in logic. The *fundamental laws of set algebra* are:

1. Associative Laws

$$(A \cap B) \cap C = A \cap (B \cap C), \quad (5.43) \qquad (A \cup B) \cup C = A \cup (B \cup C). \quad (5.44)$$

2. Commutative Laws

$$A \cap B = B \cap A, \quad (5.45) \qquad A \cup B = B \cup A. \quad (5.46)$$

3. Distributive Laws

$$(A \cup B) \cap C = (A \cap C) \cup (B \cap C), \quad (5.47) \qquad (A \cap B) \cup C = (A \cup C) \cap (B \cup C). \quad (5.48)$$

4. Absorption Laws

$$A \cap (A \cup B) = A, \quad (5.49) \qquad A \cup (A \cap B) = A. \quad (5.50)$$

5. Idempotence Laws

$$A \cap A = A, \quad (5.51) \qquad A \cup A = A. \quad (5.52)$$

6. De Morgan Laws

$$\overline{A \cap B} = \overline{A} \cup \overline{B}, \quad (5.53) \qquad \overline{A \cup B} = \overline{A} \cap \overline{B}. \quad (5.54)$$

7. Some Further Laws

$$A \cap \overline{A} = \emptyset, \quad (5.55) \qquad A \cup \overline{A} = M \quad (M \text{ fundamental set}), \quad (5.56)$$

$$A \cap M = A, \quad (5.57) \qquad A \cup \emptyset = A, \quad (5.58)$$

$$A \cap \emptyset = \emptyset, \quad (5.59) \qquad A \cup M = M, \quad (5.60)$$

$$\overline{M} = \emptyset, \quad (5.61) \qquad \overline{\emptyset} = M. \quad (5.62)$$

$$\overline{\overline{A}} = A. \quad (5.63)$$

This table can also be obtained from the fundamental laws of propositional calculus (see 5.1.1, p. 323) using the following substitutions: $\wedge$ by $\cap$, $\vee$ by $\cup$, T by M, and F by $\emptyset$. This coincidence is not accidental; it will be discussed in 5.7, p. 395.

4. Further Set Operations

In addition to the operations defined above there are defined some further operations between two sets A and B: the *difference set* or *difference* $A \setminus B$, the *symmetric difference* $A \triangle B$ and the *Cartesian product* $A \times B$.

1. Difference of Two Sets The set of the elements of A, not belonging to B is the *difference set* or *difference* of A and B:

$$A \setminus B = \{x \mid x \in A \wedge x \notin B\}. \quad (5.64a)$$

If A is defined by the property E_1 and B by the property E_2, then $A \setminus B$ contains the elements having the property E_1 but not having property E_2.

In **Fig. 5.5** the difference is represented by the shaded region.

■ $\{1, 2, 3, 4\} \setminus \{3, 4, 5\} = \{1, 2\}$.

2. Symmetric Difference of Two Sets The symmetric difference $A \triangle B$ is the set of all elements belonging to exactly one of the sets A and B:

$$A \triangle B = \{x \mid (x \in A \wedge x \notin B) \vee (x \in B \wedge x \notin A)\}. \quad (5.64b)$$

It follows from the definition that

$$A \triangle B = (A \setminus B) \cup (B \setminus A) = (A \cup B) \setminus (A \cap B), \quad (5.64c)$$

i.e., the symmetric difference contains the elements which have exactly one of the defining properties E_1 (for A) and E_2 (for B).

In **Fig. 5.6** the symmetric difference is represented by the shaded region.

■ $\{1,2,3,4\} \triangle \{3,4,5\} = \{1,2,5\}$.

3. Cartesian Product of Two Sets The *Cartesian product* of two sets $A \times B$ is defined by

$$A \times B = \{(a,b) \mid a \in A \wedge b \in B\}. \tag{5.65a}$$

The elements (a,b) of $A \times B$ are called *ordered pairs* and they are characterized by

$$(a,b) = (c,d) \Leftrightarrow a = c \wedge b = d. \tag{5.65b}$$

The number of the elements of a Cartesian product of two finite sets is equal to

$$\text{card}\,(A \times B) = (\text{card}A)(\text{card}B). \tag{5.65c}$$

■ **A:** For $A = \{1,2,3\}$ and $B = \{2,3\}$ one gets $A \times B = \{(1,2),(1,3),(2,2),(2,3),(3,2),(3,3)\}$ and $B \times A = \{(2,1),(2,2),(2,3),(3,1),(3,2),(3,3)\}$ with $\text{card}A = 3$, $\text{card}B = 2$, $\text{card}(A \times B) = \text{card}(B \times A) = 6$.

■ **B:** Every point of the x,y plane can be defined with the Cartesian product $\mathbb{R} \times \mathbb{R}$ ($\mathbb{R}$ is the set of real numbers). The set of the coordinates x, y is represented by $\mathbb{R} \times \mathbb{R}$, so:

$$\mathbb{R}^2 = \mathbb{R} \times \mathbb{R} = \{(x,y) \mid x \in \mathbb{R}, y \in \mathbb{R}\}.$$

4. Cartesian Product of *n* Sets

From n elements, by fixing an order of sequence (first element, second element, ..., n-th element) an ordered n tuple is defined. If $a_i \in A_i$ $(i = 1,2,\ldots,n)$ are the elements, the n tuple is denoted by $(a_1, a_2, \ldots, a_n)$, where a_i is called the i-th component.

For $n = 3,4,5$ these n tuples are called *triples, quadruples*, and *quintuples*.

The Cartesian product of n terms $A_1 \times A_2 \times \cdots \times A_n$ is the set of all ordered n tuples $(a_1, a_2, \ldots, a_n)$ with $a_i \in A_i$:

$$A_1 \times \ldots \times A_n = \{(a_1, \ldots, a_n) \mid a_i \in A_i \;\; (i = 1, \ldots, n)\}. \tag{5.66a}$$

If every A_i is a finite set, the number of ordered n tuples is

$$\text{card}(A_1 \times A_2 \times \cdots \times A_n) = \text{card}A_1\,\text{card}A_2 \cdots \text{card}A_n. \tag{5.66b}$$

Remark: The n times Cartesian product of a set A with itself is denoted by A^n.

5.2.3 Relations and Mappings

1. *n* ary Relations

Relations define correspondences between the elements of one or different sets. An *n ary relation* or *n-place relation* R between the sets $A_1, \ldots, A_n$ is a subset of the Cartesian product of these sets, i.e., $R \subseteq A_1 \times \ldots \times A_n$. If the sets A_i, $i = 1, \ldots, n$, are all the same set A, then $R \subseteq A^n$ holds and it is called an n ary relation in the set A.

2. Binary Relations

1. Notion of Binary Relations of a Set The two-place *(binary)* relations in a set have special importance.

In the case of a binary relation the notation aRb is also very common instead of $(a,b) \in R$.

■ As an example, the divisibility relation in the set $A = \{1,2,3,4\}$ is considered, i.e., the binary relation

$$T = \{(a,b) \mid a, b \in A \wedge \; a \text{ is a divisor of } b\} \tag{5.67a}$$

$$= \{(1,1),(1,2),(1,3),(1,4),(2,2),(2,4),(3,3),(4,4)\}. \tag{5.67b}$$

2. Arrow Diagram or Mapping Function Finite binary relations R in a set A can be represented by *arrow functions* or *arrow diagrams* or by *relation matrices*. The elements of A are represented as points of the plane and an arrow goes from a to b if aRb holds.

Fig. 5.7 shows the arrow diagram of the relation T in $A = \{1,2,3,4\}$.

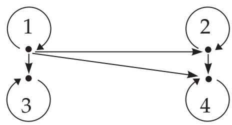


Figure 5.7

	1	2	3	4
1	1	1	1	1
2	0	1	0	1
3	0	0	1	0
4	0	0	0	1

Scheme: Relation matrix

3. Relation Matrix The elements of A are used as row and column entries of a matrix (see 4.1.1, **1.**, p. 269). At the intersection point of the row of $a \in A$ with the column of $b \in B$ there is an entry 1 if aRb holds, otherwise there is an entry 0. The above scheme shows the relation matrix for T in $A = \{1, 2, 3, 4\}$.

3. Relation Product, Inverse Relation

Relations are special sets, so the usual set operations (see 5.2.2, p. 328) can be performed between relations. Besides them, for binary relations, the *relation product* and the *inverse relation* also have special importance.
Let $R \subseteq A \times B$ and $S \subseteq B \times C$ be two binary relations. The product $R \circ S$ of the relations R, S is defined by

$$R \circ S = \{(a, c) \mid \exists\, b\, (b \in B \wedge aRb \wedge bSc)\}. \tag{5.68}$$

The relation product is associative, but not commutative.

The inverse relation R^{-1} of a relation R is defined by

$$R^{-1} = \{(b, a) \mid (a, b) \in R\}. \tag{5.69}$$

For binary relations in a set A the following relations are valid:

$$(R \cup S) \circ T = (R \circ T) \cup (S \circ T), \quad \text{(5.70a)} \qquad (R \cap S) \circ T \subseteq (R \circ T) \cap (S \circ T), \quad \text{(5.70b)}$$

$$(R \cup S)^{-1} = R^{-1} \cup S^{-1}, \quad \text{(5.70c)} \qquad (R \cap S)^{-1} = R^{-1} \cap S^{-1}, \quad \text{(5.70d)}$$

$$(R \circ S)^{-1} = S^{-1} \circ R^{-1}. \quad \text{(5.70e)}$$

4. Properties of Binary Relations

A binary relation in a set A can have special important properties:
R is called

$$\begin{aligned} &reflexive, && \text{if } \forall\, a \in A\; aRa, && \text{(5.71a)}\\ &irreflexive, && \text{if } \forall\, a \in A\; \neg aRa, && \text{(5.71b)}\\ &symmetric, && \text{if } \forall\, a, b \in A\; (aRb \Rightarrow bRa), && \text{(5.71c)}\\ &antisymmetric, && \text{if } \forall\, a, b \in A\; (aRb \wedge bRa \Rightarrow a = b), && \text{(5.71d)}\\ &transitive, && \text{if } \forall\, a, b, c \in A\; (aRb \wedge bRc \Rightarrow aRc), && \text{(5.71e)}\\ &linear, && \text{if } \forall\, a, b \in A\; (aRb \vee bRa). && \text{(5.71f)} \end{aligned}$$

These relations can also be described by the relation product. For instance: a binary relation is transitive if $R \circ R \subseteq R$ holds. Especially interesting is the *transitive closure* $\mathrm{tra}(R)$ of a relation R. It is the smallest (with respect to the subset relation) transitive relation which contains R. In fact

$$\mathrm{tra}(R) = \bigcup_{n \geq 1} R^n = R^1 \cup R^2 \cup R^3 \cup \cdots, \tag{5.72}$$

where R^n is the n times relation product of R with itself.

■ Let a binary relation R on the set $\{1, 2, 3, 4, 5\}$ be given by its relation matrix M:

M	1	2	3	4	5
1	1	0	0	1	0
2	0	0	0	1	0
3	0	0	1	0	1
4	0	1	0	0	1
5	0	1	0	0	0

M^2	1	2	3	4	5
1	1	1	0	1	1
2	0	1	0	0	1
3	0	1	1	0	1
4	0	1	0	1	0
5	0	0	0	1	0

M^3	1	2	3	4	5
1	1	1	0	1	1
2	0	1	0	1	0
3	0	1	1	1	1
4	0	1	0	1	1
5	0	1	0	0	1

Calculating M^2 by matrix multiplication where the values 0 and 1 are treated as truth values and instead of multiplication and addition one performs the logical operations conjunction and disjunction, then, M^2 is the relation matrix belonging to R^2. The relation matrices of R^3, R^4 etc. can be calculated similarly.

$M \vee M^2 \vee M^3$	1	2	3	4	5
1	1	1	0	1	1
2	0	1	0	1	1
3	0	1	1	1	1
4	0	1	0	1	1
5	0	1	0	1	1

The relation matrix of $R \cup R^2 \cup R^3$ (the matrix on the left) can be get by calculating the disjunction elementwise of the matrices M, M^2 and M^3. Since the higher powers of M contains no new 1-s, this matrix already coincides with the relation matrix of $\mathrm{tra}(R)$.

The relation matrix and relation product have important applications in search of path length in graph theory (see 5.8.2.1, p. 404).

In the case of finite binary relations, one can easily recognize the above properties from the arrow diagrams or from the relation matrices. For instance one can recognize the reflexivity from "self-loops" in the arrow diagram, and from the main diagonal elements 1 in the relation matrix. Symmetry is obvious in the arrow diagram if to every arrow there belongs another one in the opposite direction, or if the relation matrix is a symmetric matrix (see 5.2.3, **2.**, p. 331). Easy to see from the arrow diagram or from the relation matrix that the divisibility T is a reflexive but not symmetric relation.

5. Mappings

A *mapping* or *function* f (see 2.1.1.1, p. 48) from a set A to a set B with the notation $f\colon A \to B$ is a rule to assign to every element $a \in A$ exactly one element $b \in B$, which is called $f(a)$.
A mapping f can be considered as a subset of $A \times B$ and so as a binary relation:

$$f = \{(a, f(a)) | a \in A\} \subseteq A \times B\,. \tag{5.73}$$

a) f is called a *injective* or *one to one* mapping, if to every $b \in B$ at most one $a \in A$ with $f(a) = b$ exists.
b) f is called a *surjective mapping* from A *to* B, if to every $b \in B$ at least one $a \in A$ with $f(a) = b$ exists.
c) f is called *bijective*, if f is both injective and surjective.

If A and B are finite sets, between which exists a bijective mapping, then A and B possess the same number of elements (see also 5.2.5, p. 335).
For a bijective mapping $f\colon A \to B$ exists the inverse relation $f^{-1}\colon B \to A$, the so-called *inverse mapping* of f.
The relation product of mappings is used for the one after the other composition of mappings: If f: $A \to B$ and g: $B \to C$ are mappings, then $f \circ g$ is also a mapping from A to C, and is defined by

$$(f \circ g)(a) = g(f(a))\,. \tag{5.74}$$

Remark: Be careful with the order of f and g in this equation (it is treated differently in the literature!).

5.2.4 Equivalence and Order Relations

The most important classes of binary relations with respect to a set A are the equivalence and order relations.

1. Equivalence Relations

A binary relation R with respect to a set A is called an *equivalence relation* if R is reflexive, symmetric, and transitive. For aRb also the notations $a \sim_R b$ or $a \sim b$ are used, if the equivalence relation R is already known, in words a is equivalent to b (with respect to R).

Examples of Equivalence Relations:

■ **A:** $A = \mathbb{Z}$, $m \in \mathbb{N} \setminus \{0\}$. $a \sim_R b$ holds exactly if a and b have the same remainder when divided by m (they are congruent modulo m).

■ **B:** Equality relation in different domains, e.g., in the set $\mathbb{Q}$ of rational numbers: $\frac{p_1}{q_1} = \frac{p_2}{q_2} \Leftrightarrow p_1 q_2 = p_2 q_1$ (p_1, p_2, q_1, q_2 integer; $q_1, q_2 \neq 0$), where the first equality sign defines an equality in $\mathbb{Q}$, while the second one denotes an equality in $\mathbb{Z}$.

■ **C:** Similarity or congruence of geometric figures.

■ **D:** Logical equivalence of expressions of propositional calculus (see 5.1.1, **6.**, p. 324).

2. Equivalence Classes, Partitions

1. Equivalence Classes An equivalence relation in a set A defines a partition of A into non-empty pairwise disjoint subsets, into *equivalence classes*.

$$[a]_R := \{b \mid b \in A \wedge a \sim_R b\} \tag{5.75}$$

is called an equivalence class of a with respect to R. For equivalence classes the following is valid:

$$[a]_R \neq \emptyset, \quad a \sim_R b \Leftrightarrow [a]_R = [b]_R\,, \quad \text{and} \quad a \not\sim_R b \Leftrightarrow [a]_R \cap [b]_R = \emptyset. \tag{5.76}$$

These equivalence classes form a new set, the *quotient set* A/R:

$$A/R = \{[a]_R \mid a \in A\}. \tag{5.77}$$

A subset $Z \subseteq \mathbb{P}(A)$ of the power set $\mathbb{P}(A)$ is called a *partition* of A if

$$\emptyset \notin Z, \quad X, Y \in Z \wedge X \neq Y \Rightarrow X \cap Y = \emptyset\,, \quad \bigcup_{X \in Z} X = A. \tag{5.78}$$

2. Decomposition Theorem Every equivalence relation R in a set A defines a partition Z of A, namely $Z = A/R$. Conversely, every partition Z of a set A defines an equivalence relation R in A:

$$a \sim_R b \Leftrightarrow \exists X \in Z\,(a \in X \wedge b \in X). \tag{5.79}$$

An equivalence relation in a set A can be considered as a generalization of the equality, where " insignificant " properties of the elements of A are neglected, and the elements, which do not differ with respect to a certain property, belong to the same equivalence class.

3. Ordering Relations

A binary relation R in a set A is called a *partial ordering* if R is reflexive, antisymmetric, and transitive. If in addition R is linear, then R is called a *linear ordering* or a *chain*. The set A is called ordered or linearly ordered by R. In a linearly ordered set any two elements are comparable. Instead of aRb also the notation $a \leq_R b$ or $a \leq b$ is used, if the ordering relation R is known from the problem.

Examples of Ordering Relations:

■ **A:** The sets of numbers $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$ are completely ordered by the usual $\leq$ relation.

■ **B:** The subset relation is also an ordering, but only a partial ordering.

■ **C:** The *lexicographical order* of the English words is a chain.

Remark: If $Z = \{A, B\}$ is a partition of $\mathbb{Q}$ with the property $a \in A \wedge b \in B \Rightarrow a < b$, then (A, B) is called a *Dedekind cut*. If neither A has a greatest element nor B has a smallest element, so an irrational number is uniquely determined by this cut. Besides the nest of intervals (see 1.1.1.2, p. 2) the notion of Dedekind cuts is another way to introduce irrational numbers.

4. Hasse Diagram

Finite ordered sets can be represented by the *Hasse diagram*: Let an ordering relation $\leq$ be given on a finite set A. The elements of A are represented as points of the plane, where the point $b \in A$ is placed

above the point $a \in A$ if $a < b$ holds. If there is no $c \in A$ for which $a < c < b$, one says a and b are *neighbors* or *consecutive members*. Then one connects a and b by a line segment.

A Hasse diagram is a "simplified" arrow diagram, where all the loops, arrowheads, and the arrows following from the transitivity of the relation are eliminated. The arrow diagram of the divisibility relation T of the set $A = \{1, 2, 3, 4\}$ is given in **Fig. 5.7**. T also denotes an ordering relation, which is represented by the Hasse diagram in **Fig. 5.8**.

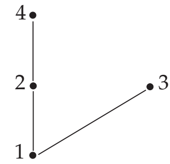


Figure 5.8

5.2.5 Cardinality of Sets

In 5.2.1, p. 327 the number of elements of a finite set was called the cardinality of the set. This notion of cardinality can be extended to infinite sets.

1. Cardinal Numbers

Two sets A and B are called *equinumerous* if there is a bijective mapping between them. To every set A a *cardinal number* $|A|$ or card A is assigned, so that equinumerous sets have the same cardinal number. A set and its power set are never equinumerous, so no " greatest " cardinal number exists.

2. Infinite Sets

Infinite sets can be characterized by the property that they have proper subsets equinumerous to the set itself. The "smallest" infinite cardinal number is the cardinal number of the set $\mathbb{N}$ of the natural numbers. This is denoted by $\aleph_0$ (aleph 0).

A set is called *enumerable* or *countable* if it is equinumerous to $\mathbb{N}$. This means that its elements can be enumerated or written as an infinite sequence $a_1, a_2, \ldots$.

A set is called *non-countable* if it is infinite but it is not equinumerous to $\mathbb{N}$. Consequently every infinite set which is not enumerable is non-countable.

■ **A:** The set $\mathbb{Z}$ of integers and the set $\mathbb{Q}$ of the rational numbers are countable sets.

■ **B:** The set $\mathbb{R}$ of the real numbers and the set $\mathbb{C}$ of the complex numbers are non-countable sets. These sets are equinumerous to $\mathbb{P}(\mathbb{N})$, the power set of the natural numbers, and their cardinality is called the *continuum*.

5.3 Classical Algebraic Structures

5.3.1 Operations

1. n ary Operations

The notion of structure has a central role in mathematics and its applications. Next to investigate are algebraic structures, i.e., sets on which operations are defined. An *n ary operation* φ on a set A is a mapping φ: $A^n \to A$, which assigns an element of A to every n tuple of elements of A.

2. Properties of Binary Operations

Especially important is the case $n = 2$, which is called a *binary operation*, e.g., addition and multiplication of numbers or matrices, or union and intersection of sets. A binary operation can be considered as a mapping $* : A \times A \to A$, where instead of the notation "$*(a, b)$" in this chapter mostly the *infix form* "$a * b$" will be used. A binary operation $*$ in A is called *associative* if

$$(a * b) * c = a * (b * c), \tag{5.80}$$

and *commutative* if

$$a * b = b * a \tag{5.81}$$

holds for every $a, b, c \in A$.
An element $e \in A$ is called a *neutral element* with respect to a binary operation $*$ in A if

$$a * e = e * a = a \quad \text{holds for every} \quad a \in A. \tag{5.82}$$

3. Exterior Operations

Sometimes exterior operations are to be considered. That are the mappings from $K \times A$ to K, where K is an "exterior" and mostly already structured set (see 5.3.8, p. 365).

5.3.2 Semigroups

The most frequently occurring algebraic structures have their own names. A set H having one associative binary operation $*$, is called a *semigroup*. The notation: is $H = (H, *)$.

Examples of Semigroups:

■ **A:** Number domains with respect to addition or multiplication.

■ **B:** Power sets with respect to union or intersection.

■ **C:** Matrices with respect to addition or multiplication.

■ **D:** The set A^* of all " words " (strings) over an " alphabet " A with respect to concatenation (*free semigroup*).

Remark: Except for multiplication of matrices and concatenation of words, all operations in these examples are also commutative; in this case one talks about a commutative semigroup.

5.3.3 Groups

5.3.3.1 Definition and Basic Properties

1. Definition, Abelian Group

A set G with a binary operation $*$ is called a *group* if

- $*$ is associative,
- $*$ has a neutral element e, and for every element $a \in G$ there exists an *inverse element* a^{-1} such that

$$a * a^{-1} = a^{-1} * a = e. \tag{5.83}$$

A group is a special semigroup.

The neutral element of a group is unique, i.e., there exists only one. Furthermore, every element of the group has exactly one inverse. If the operation $*$ is commutative, then the group is called an *Abelian group*. If the group operation is written as addition, $+$, then the neutral element is denoted by 0 and the inverse of an element a by $-a$.

The number of elements of a finite group is called the *order of the group* (see 5.3.3.2,**3.**, p. 338).

Examples of Groups:

■ **A:** The number of domains (except $\mathbb{N}$) with respect to addition.

■ **B:** $\mathbb{Q} \setminus \{0\}$, $\mathbb{R} \setminus \{0\}$, and $\mathbb{C} \setminus \{0\}$ with respect to multiplication.

■ **C:** $S_M := \{f : M \to M \wedge f \text{bijective}\}$ with respect to composition of mappings. This group is called symmetric. If M is finite having n elements, then S_n is written instead of S_M. S_n has $n!$ elements.

The symmetric group S_n and its subgroups are called *permutation groups*. So, the dieder groups D_n are permutation groups and subgroups of S_n.

■ **D:** The set D_n of all covering transformations of a regular n-gon in the plane is considered. Here a *covering transformation* is the transition between two symmetric positions of the n-gon, i.e., the moving of the n-gon into a superposable position. Denoting by d a rotation by the angle $2\pi/n$ and by σ the reflection with respect to an axis, then D_n has $2n$ elements:

$$D_n = \{e, d, d^2, \ldots, d^{n-1}, \sigma, d\sigma, \ldots, d^{n-1}\sigma\}.$$

With respect to the composition of mappings D_n is a group, the *dihedral group*. Here the equalities $d^n = \sigma^2 = e$ and $\sigma d = d^{n-1}\sigma$ hold.

■ **E:** All the regular matrices (see 4.1.4, p. 272) over the real or complex numbers with respect to multiplication.

Remark: Matrices have a very important role in applications, especially in representation of linear transformations. Linear transformations can be classified by matrix groups.

2. Group Tables or Cayley's Tables

For the representation of finite groups Cayley's tables or group tables are used: The elements of the group are denoted at the row and column headings. The element $a * b$ is the intersection of the row of the element a and the column of the element b.

■ If $M = \{1, 2, 3\}$, then the symmetric group S_M is also denoted by S_3. S_3 consists of all the bijective mappings (permutations) of the set $\{1, 2, 3\}$ and consequently it has $3! = 6$ elements (see 16.1.1, p. 805). Permutations are mostly represented in two rows, where in the first row there are the elements of M and under each of them there is its image. So one gets the six elements of S_3 as follows:

$$\begin{aligned} \varepsilon &= \begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix}, \quad p_1 = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix}, \quad p_2 = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix}, \\ p_3 &= \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix}, \quad p_4 = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix}, \quad p_5 = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix}. \end{aligned} \tag{5.84}$$

With the successive application of these mappings (binary operations) the following group table is obtained for S_3:

$\circ$	ε	p_1	p_2	p_3	p_4	p_5
ε	ε	p_1	p_2	p_3	p_4	p_5
p_1	p_1	ε	p_5	p_4	p_3	p_2
p_2	p_2	p_4	ε	p_5	p_1	p_3
p_3	p_3	p_5	p_4	ε	p_2	p_1
p_4	p_4	p_2	p_3	p_1	p_5	ε
p_5	p_5	p_3	p_1	p_2	ε	p_4

(5.85)

• From the group table it can be seen that the identity permutation ε is the neutral element of the group.

• In the group table every element appears exactly once in every row and in every column.

• It is easy to recognize the inverse of any group element in the table, i.e., the inverse of p_4 in S_3 is the permutation p_5, because at the intersection of the row of p_4 with the column of p_5 is the neutral element ε.

• If the group operation is commutative (Abelian group), then the table is symmetric with respect to the "main diagonal"; S_3 is not commutative, since, e.g., $p_1 \circ p_2 \neq p_2 \circ p_1$.

• The associative property cannot be easily recognized from the table.

5.3.3.2 Subgroups and Direct Products

1. Subgroups

Let $G = (G, *)$ be a group and $U \subseteq G$. If U is also a group with respect to $*$, then $U = (U, *)$ is called a *subgroup* of G.

A non-empty subset U of a group $(G, *)$ is a subgroup of G if and only if for every $a, b \in U$, the elements $a * b$ and a^{-1} are also in U (*subgroup criterion*).

1. Cyclic Subgroups The group G itself and $E = \{e\}$ are subgroups of G, the so-called *trivial subgroups*. Furthermore, a subgroup corresponds to every element $a \in G$, the so-called *cyclic subgroup* generated by a:

$$< a > = \{\ldots, a^{-2}, a^{-1}, e, a, a^2, \ldots\}. \tag{5.86}$$

If the group operation is addition, then one writes the integer multiple ka as a shorthand notation of the k times addition of a with itself instead of the power a^k, i.e., as a shorthand notation of the k times operation of a by itself,

$$< a > = \{\ldots, (-2)a, -a, 0, a, 2a, \ldots\}. \tag{5.87}$$

Here $< a >$ is the smallest subgroup of G containing a. If $< a > = G$ holds for an element a of G, then G is called cyclic.

There are infinite cyclic groups, e.g., $\mathbb{Z}$ with respect to addition, and finite cyclic groups, e.g., the set $\mathbb{Z}_m$ the residue class modulo m with residue class addition (see 5.4.3, **3.**, p. 377).

■ If the number of elements of a finite G group is a prime, then G is always cyclic.

2. Generalization The notion of cyclic groups can be generalized as follows: If M is a non-empty subset of a group G, then the subgroup of G whose elements can be written in the form of a product

of finitely many elements of M and their inverses, is denoted by $< M >$. The subset M is called the *system of generators* of $< M >$. If M contains only one element, then $< M >$ is cyclic.

3. Order of a Group, Left and Right Cosets In group theory the number of elements of a finite group is denoted by ord G. If the cyclic subgroup $< a >$ generated by one element a is finite, then this order is also called the *order of the element* a, i.e., $\operatorname{ord} < a > = \operatorname{ord} a$.
If U is a subgroup of a group $(G, *)$ and $a \in G$, then the subsets

$$aU := \{a * u | u \in U\} \quad \text{and} \quad Ua := \{u * a | u \in U\} \tag{5.88}$$

of G are called *left co-sets* and *right co-sets* of U in G. The left or right co-sets form a partition of G, respectively (see 5.2.4, **2.**, p. 334).

All the left or right co-sets of a subgroup U in a group G have the same number of elements, namely ord$U$. From this it follows that the number of left co-sets is equal to the number of right co-sets. This number is called the *index* of U in G. The Lagrange theorem follows from these facts.

4. Lagrange Theorem The order of a subgroup is a divisor of the order of the group.
In general it is difficult to determine all the subgroups of a group. In the case of finite groups the Lagrange theorem as a necessary condition for the existence of a subgroup is useful.

2. Normal Subgroup or Invariant Subgroup

For a subgroup U, in general, aU is different from Ua (however $|aU| = |Ua|$ is valid). If $aU = Ua$ for all $a \in G$ holds, then U is called a *normal subgroup* or *invariant subgroup* of G. These special subgroups are the basis of forming factor groups (see 5.3.3.3, **3.**, p. 339).
In Abelian groups, obviously, every subgroup is a normal subgroup.

Examples of Subgroups and Normal Subgroups:

■ **A:** $\mathbb{R} \setminus \{0\}$, $\mathbb{Q} \setminus \{0\}$ form subgroups of $\mathbb{C} \setminus \{0\}$ with respect to multiplication.

■ **B:** The even integers form a subgroup of $\mathbb{Z}$ with respect to addition.

■ **C:** Subgroups of S_3: According to the Lagrange theorem the group S_3 having six elements can have subgroups only with two or three elements (besides the trivial subgroups). In fact, the group S_3 has the following subgroups: $E = \{\varepsilon\}$, $U_1 = \{\varepsilon, p_1\}$, $U_2 = \{\varepsilon, p_2\}$, $U_3 = \{\varepsilon, p_3\}$, $U_4 = \{\varepsilon, p_4, p_5\}$, S_3.
The non-trivial subgroups U_1, U_2, U_3, and U_4 are cyclic, since the numbers of their elements are primes. But the group S_3 is not cyclic. The group S_3 has only U_4 as a normal subgroup, except the trivial normal subgroups.
Anyway, every subgroup U of a group G with $|U| = |G|/2$ is a normal subgroup of G.
Every symmetric group S_{M} and their subgroups are called *permutation groups*.

■ **D:** Special subgroups of the group $GL(n)$ of all regular matrices of type (n, n) with respect to matrix multiplication:

$SL(n)$ group of all matrices A with determinant 1,
$O(n)$ group of all orthogonal matrices,
$SO(n)$ group of all orthogonal matrices with determinant 1.

The group $SL(n)$ is a normal subgroup of $GL(n)$ (see 5.3.3.3, **3.**, p. 339) and $SO(n)$ is a normal subgroup of $O(n)$.

■ **E:** As subgroups of all complex matrices of type (n, n) (see 4.1.4, p. 272):

$U(n)$ group of all unitary matrices,
$SU(n)$ group of all unitary matrices with determinant 1.

3. Direct Product

1. Definition Suppose A and B are groups, whose group operation (e.g., addition or multiplication) is denoted by $\cdot$. In the Cartesian product (see 5.2.2, **4.**, p. 331) $A \times B$ (5.65a) an operation $*$ can be introduced in the following way:

$$(a_1, b_1) * (a_2, b_2) = (a_1 \cdot a_2, b_1 \cdot b_2). \tag{5.89a}$$

$A \times B$ becomes a group with this operation and it is called the *direct product* of A and B.
(e, e) denotes the unit element of $A \times B$, (a^{-1}, b^{-1}) is the inverse element of (a, b).

For finite groups A, B

$$\operatorname{ord}(A \times B) = \operatorname{ord} A \cdot \operatorname{ord} B \tag{5.89b}$$

holds. The groups $A' := \{(a, e)|a \in A\}$ and $B' := \{(e, b)|b \in B\}$ are normal subsets of $A \times B$ isomorphic to A and B, respectively.
The direct product of Abelian groups is again an Abelian group.
The direct product of two cyclic groups A, B is cyclic if and only if the greatest common divisor of the orders of the groups is equal to 1.

■ **A:** With $Z_2 = \{e, a\}$ and $Z_3 = \{e, b, b^2\}$, the direct product $Z_2 \times Z_3 = \{(e, e), (e, b), (e, b^2), (a, e), (a, b), (a, b^2)\}$, is a group isomorphic to Z_6 (see 5.3.3.3, **2.**, p. 339) generated by (a, b).

■ **B:** On the other hand $Z_2 \times Z_2 = \{(e, e), (e, b), (a, e), (a, b)\}$ is not cyclic. This group has order 4 and it is also called Klein's four group, and it describes the covering operations of a rectangle.

2. Fundamental Theorem of Abelian Groups Because the direct product is a construction which enables to make "larger" groups from "smaller" groups, the question can be reversed: When is it possible to consider a larger group G as a direct product of smaller groups A, B, i.e., when will G be isomorphic to $A \times B$? For Abelian groups, there exists the so-called *fundamental theorem*:
Every finite Abelian group can be represented as a direct product of cyclic groups with orders of prime powers.

5.3.3.3 Mappings Between Groups

1. Homomorphism and Isomorphism

1. Group Homomorphism Between algebraic structures, not arbitrary mappings, but only "structure keeping" mappings are considered:
Let $G_1 = (G_1, *)$ and $G_2 = (G_2, \circ)$ are two groups. A mapping h: $G_1 \to G_2$ is called a *group homomorphism*, if for all $a, b \in G_1$ holds:

$$h(a * b) = h(a) \circ h(b) \quad (\text{"image of product = product of images"}) \tag{5.90}$$

■ As an example, consider the multiplication law for determinants (see 4.2.2, **7.**, p. 279):

$$\det(AB) = (\det A)(\det B). \tag{5.91}$$

Here on the right-hand side there is the product of non-zero numbers, on the left-hand side there is the product of regular matrices.
If h: $G_1 \to G_2$ is a group homomorphism, then the set of elements of G_1, whose image is the neutral element of G_2, is called the *kernel* of h, and it is denoted by $\ker h$. The kernel of h is a normal subgroup of G_1.

2. Group Isomorphism If a group homomorphism h is also bijective, then h is called a *group isomorphism*, and the groups G_1 and G_2 are called *isomorphic* to each other (notation: $G_1 \cong G_2$). Then $\ker h = E$ is valid.
Isomorphic groups have the same structure, i.e., they differ only by the notation of their elements.

■ The symmetric group S_3 and the dihedral group D_3 are isomorphic groups of order 6 and describe the covering mappings of an equilateral triangle.

2. Cayley's Theorem

The Cayley theorem says that *every* group can be interpreted as a permutation group (see 5.3.3.2, **2.**, p. 338):
Every group is isomorphic to a permutation group.
The permutation group P, whose elements are the permutations π_g $(g \in G)$ mapping a to $G, *g$, is a subgroup of S_G isomorphic to $(G, *)$.

3. Homomorphism Theorem for Groups

The set of co-sets of a normal subgroup N in a group G is also a group with respect to the operation

$$aN \circ bN = abN. \tag{5.92}$$

It is called the *factor group* of G with respect to N, and it is denoted by G/N.

The following theorem gives the correspondence between homomorphic images and factor groups of a group, because of what it is called the homomorphism theorem for groups:
A group homomorphism h: $G_1 \to G_2$ defines a normal subgroup of G_1, namely $\ker h = \{a \in G_1 | h(a) = e\}$. The factor group $G_1/\ker h$ is isomorphic to the homomorphic image $h(G_1) = \{h(a) | a \in G_1\}$. Conversely, every normal subgroup N of G_1 defines a homomorphic mapping nat_N: $G_1 \to G_1/N$ with $nat_N(a) = aN$. This mapping nat_N is called a *natural homomorphism*.

■ Since the determinant construction det: $GL(n) \to \mathbb{R} \setminus \{0\}$ is a group homomorphism with kernel $SL(n)$, $SL(n)$ is a normal subgroup of $GL(n)$ and (according to the homomorphism theorem): $GL(n)/SL(n)$ is isomorphic to the multiplicative group $\mathbb{R}\setminus\{0\}$ of real numbers (for notation see 5.3.3.2, **2.**, p. 338).

5.3.4 Group Representations

5.3.4.1 Definitions

1. Representation

A *representation* $D(G)$ of the *group* G is a map (homomorphism) of G onto the group of non-singular linear transformations D on an n-dimensional (real or complex) vector space V_n:

$$D(G) : a \to D(a), \quad a \in G. \tag{5.93}$$

The vector space V_n is called the *representation space*; n is the dimension of the representation (see also 12.1.3, **2.**, p. 657). Introducing the basis $\{\underline{\mathbf{e}}_i\}$ $(i = 1, 2, \ldots, n)$ in V_n every vector $\underline{\mathbf{x}}$ can be written as a linear combination of the basis vectors:

$$\underline{\mathbf{x}} = \sum_{i=1}^{n} x_i \underline{\mathbf{e}}_i, \quad \underline{\mathbf{x}} \in \mathrm{V}_n. \tag{5.94}$$

The action of the linear transformation $D(a)$, $a \in G$, on $\underline{\mathbf{x}}$ can be defined by the quadratic matrix $\mathbf{D}(a) = (D_{ik}(a))$ $(i, k = 1, 2, \ldots, n)$, which provides the coordinates of the transformed vector $\underline{\mathbf{x}}'$ within the basis $\{\underline{\mathbf{e}}_i\}$:

$$\underline{\mathbf{x}}' = \mathbf{D}(a)\underline{\mathbf{x}} = \sum_{i=1}^{n} x_i' \underline{\mathbf{e}}_i, \qquad x_i' = \sum_{k=1}^{n} D_{ik}(a) x_k. \tag{5.95}$$

This transformation may also be considered as a transformation of the basis $\{\underline{\mathbf{e}}_i\} \to \{\underline{\mathbf{e}}_i'\}$:

$$\underline{\mathbf{e}}_i' = \underline{\mathbf{e}}_i \mathbf{D}(a) = \sum_{k=1}^{n} D_{ki}(a) \underline{\mathbf{e}}_k. \tag{5.96}$$

Thus, every element a of the group is assigned to the *representation matrix* $\mathbf{D} = (D_{ik}(a))$:

$$D(G) : a \to \mathbf{D} = (D_{ik}(a)) \quad (i, k = 1, 2, \ldots, n), a \in G. \tag{5.97}$$

The representation matrix depends on the choice of basis.

■ **A: Abelian Point Group C_n.** A regular polygon (see 3.1.5, p. 138) with n sides has a symmetry such that rotating it around an axis, which is perpendicular to the plane of the figure and goes through its center M (**Fig.5.9**) by an angle $\varphi_k = 2\pi k/n$, $k = 0, 1, \ldots, n-1$ the resulted polygon is identical to the original one (invariance of the system under certain rotations). The rotations $R_k(\varphi_k)$ form the Abelian group of points C_n. C_n is a cyclic group (see 5.3.3.2, p. 337), i.e. every element of the group can be represented as a power of a single element R_1, whose n-th power is the unit element $e = R_0$:

$$C_n = \{e, R_1, R_1^2, \ldots, R_1^{n-1}\}, \quad R_1^n = e. \tag{5.98a}$$

Let the center of an equilateral triangle ($n = 3$) be the origin (see **Fig.5.9**), then the angles of rotations and the rotations are in accor-

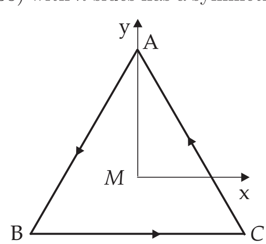


Figure 5.9

dance with (5.98b).

$$\begin{aligned} &k=0,\ \varphi_0=0 \text{ or } 2\pi,\\ &k=1,\ \varphi_1=2\pi/3,\\ &k=2,\ \varphi_2=4\pi/3. \end{aligned} \tag{5.98b}$$

$$\begin{aligned} &R_0: A\to A, B\to B, C\to C,\\ &R_1: A\to B, B\to C, C\to A,\\ &R_2: A\to C, B\to A, C\to B\,. \end{aligned} \tag{5.98c}$$

The rotations (5.98c) satisfy the relations

$$R_2 = R_1^2,\ \ R_1\cdot R_2 = R_1^3 = R_0 = e\,. \tag{5.98d}$$

They form the cyclic group C_3.
The matrix of rotation (see (3.432), p. 230)

$$\mathbf{R}(\varphi) = \begin{pmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{pmatrix} \tag{5.98e}$$

of a geometric transformation of this triangle (for rotation of this figure in a fixed coordinate system see 3.5.3.3,**3.**, p. 213) gives the representation of group C_3 if φ is substituted by the angles given in (5.98b):

$$\mathbf{D}(e) = \mathbf{R}(0) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},\ \ \mathbf{D}(R_1) = \mathbf{R}(2\pi/3) = \begin{pmatrix} -1/2 & -\sqrt{3}/2 \\ \sqrt{3}/2 & -1/2 \end{pmatrix}, \tag{5.98f}$$

$$\mathbf{D}(R_2) = \mathbf{R}(4\pi/3) = \begin{pmatrix} -1/2 & \sqrt{3}/2 \\ -\sqrt{3}/2 & -1/2 \end{pmatrix}. \tag{5.98g}$$

The same relations hold for the matrices of this representation given in (5.98f) and (5.98g) as for the group elements R_k (5.98d):

$$\mathbf{D}(R_2) = \mathbf{D}(R_1R_1) = \mathbf{D}(R_1)\mathbf{D}(R_1),\ \ \mathbf{D}(R_1)\mathbf{D}(R_2) = \mathbf{D}(e). \tag{5.98h}$$

■ **B: Dihedral Group D_3**. The equilateral triangle is invariant with respect to rotations by angle π about its bisectors (see **Fig.5.10**). These rotations correspond to reflections S_A, S_B, S_C with respect to a plane being perpendicular to the plane of the triangle and containing one of the rotation axes.

$$\begin{aligned} &S_A: \text{ Rotations } A\to A, B\to C, C\to B;\\ &S_B: \text{ Rotations } A\to C, B\to B, C\to A;\\ &S_C: \text{ Rotations } A\to B, B\to A, C\to C. \end{aligned} \tag{5.99a}$$

For the reflections there is:

$$S_\sigma S_\sigma = e \quad (\sigma = A, B, C). \tag{5.99b}$$

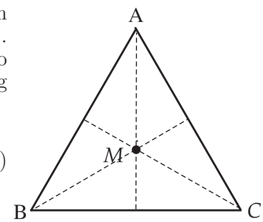


Figure 5.10

The product $S_\sigma S_\tau$ $(\sigma \neq \tau)$ results in one of the rotations R_1, R_2, e.g. using $S_A\,S_B$ for the triangle ΔABC:

$$S_A\,S_B(\Delta ABC) = S_A(\Delta CBA) = \Delta CAB = R_1(\Delta ABC)\,, \tag{5.99c}$$

consequently $S_A\,S_B = R_1$. Here S_A, S_B, S_C correspond to the outcomes on **Fig.5.10**.

The cyclic group C_3 and the reflections S_A, S_B, S_C together form the dihedral group D_3. The reflections do not form a subgroup because of (5.99c). A summary of relations is represented in group-table (5.99d).

	e	R_1	R_2	S_A	S_B	S_C
e	e	R_1	R_2	S_A	S_B	S_C
R_1	R_1	R_2	e	S_C	S_A	S_B
R_2	R_2	e	R_1	S_B	S_C	S_A
S_A	S_A	S_B	S_C	e	R_1	R_2
S_B	S_B	S_C	S_A	R_2	e	R_1
S_C	S_C	S_A	S_B	R_1	R_2	e

(5.99d)

Only the signs of the x-coordinates of points B and C are changed at reflection S_A (see **Fig.5.9**). This coordinate transformation is given by the matrix

$$\mathbf{D}(S_A) = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{5.99e}$$

The matrices representing reflections S_B and S_C can be found in the group-table (5.99d) and from the matrices of representation in (5.98f) and (5.98g)

$$\mathbf{D}(S_B) = \mathbf{D}(R_2)\mathbf{D}(S_A) = \begin{pmatrix} -1/2 & \sqrt{3}/2 \\ \sqrt{3}/2 & -1/2 \end{pmatrix} \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} -1/2 & \sqrt{3}/2 \\ \sqrt{3}/2 & -1/2 \end{pmatrix}, \tag{5.99f}$$

$$\mathbf{D}(S_C) = \mathbf{D}(R_1)\mathbf{D}(S_A) = \begin{pmatrix} -1/2 & \sqrt{3}/2 \\ \sqrt{3}/2 & -1/2 \end{pmatrix} \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1/2 & -\sqrt{3}/2 \\ -\sqrt{3}/2 & -1/2 \end{pmatrix}. \tag{5.99g}$$

Matrices (5.98f) and (5.98g) together with matrices (5.99f) and (5.99g) form a representation of the dihedral group D_3.

2. Faithful Representation

A representation is called *faithful* if $G \to D(G)$ is an isomorphism, i.e., the assignment of the element of the group to the representation matrix is a one-to-one mapping.

3. Properties of the Representations

A representation with the representation matrices $\mathbf{D}(a)$ has the following properties ($a, b \in G$, $\mathbf{I}$ unit matrix):

$$\mathbf{D}(a * b) = \mathbf{D}(a) \cdot \mathbf{D}(b), \quad \mathbf{D}(a^{-1}) = \mathbf{D}^{-1}(a), \quad \mathbf{D}(e) = \mathbf{I}. \tag{5.100}$$

5.3.4.2 Particular Representations

1. Identity Representation

Any group G has a trivial one-dimensional representation (identity representation), for which every element of the group is mapped to the unit matrix $\mathbf{I}$: $a \to \mathbf{I}$ for all $a \in G$.

2. Adjoint Representation

The representation $D^+(G)$ is called *adjoint* to $D(G)$ if the corresponding representation matrices are related by complex conjugation and reflection in the main diagonal:

$$\mathbf{D}^+(G) = \tilde{\mathbf{D}}^*(G). \tag{5.101}$$

3. Unitary Representation

For a *unitary representation* all representation matrices are unitary matrices:

$$\mathbf{D}(G) \cdot \mathbf{D}^+(G) = \mathbf{I}, \tag{5.102}$$

where $\mathbf{E}$ is the unit matrix.

4. Equivalent Representations

Two representations $D(G)$ and $D'(G)$ are called *equivalent* if for each element a of the group the corresponding representation matrices are related by the same similarity transformation with the non-singular matrix $\mathbf{T} = (T_{i,j})$:

$$\mathbf{D}'(a) = \mathbf{T}^{-1} \cdot \mathbf{D}(a) \cdot \mathbf{T}, \quad D'_{ik}(a) = \sum_{j,l=1}^{n} T_{ij}^{-1} \cdot D_{jl}(a) \cdot T_{lk}, \tag{5.103}$$

where $T^{-1}_{i,j}$ denotes the elements of the inverse matrix $\mathbf{T}^{-1}$ of $\mathbf{T}$. If such a relation does not hold two representations are called *non-equivalent*. The transition from $D(G)$ to $D'(G)$ corresponds to the transformation $T : \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\} \to \{\mathbf{e}'_1, \mathbf{e}'_2, \ldots, \mathbf{e}'_n\}$ of the basis in the representation space V_n:

$$\mathbf{e}' = \mathbf{e}\,T, \quad \mathbf{e}'_i = \sum_{k=1}^{n} T_{ki}\mathbf{e}_k \quad (i = 1, 2, \ldots, n). \tag{5.104}$$

Any representation of a finite group is equivalent to a unitary representation.

5. Character of a Group Element

In the representation $D(G)$ the *character* $\chi(a)$ of the group element a is defined as the trace of the representation matrix $\mathbf{D}(a)$ (sum of the main diagonal elements of the matrix):

$$\chi(a) = \mathrm{Tr}\,(\mathbf{D}) = \sum_{i=1}^{n} D_{ii}(a). \tag{5.105}$$

The character of the unit element e is given by the dimension n of the representation: $\chi(e) = n$. Since the trace of a matrix is invariant under similarity transformations, the group element a has the same character for equivalent representations.

■ Within the shell model of atomic or nuclear physics two out of three particles with space coordinates $\vec{\mathbf{r}}_i$ $(i = 1, 2, 3)$ can be described by the wave function $\varphi_\alpha(\vec{\mathbf{r}})$ while the third particle has the wave function $\varphi_\beta(\vec{\mathbf{r}})$ (configuration $\alpha^2\beta(\vec{\mathbf{r}})$). The wave function ψ of the system is a product of the three one-particle wave functions: $\psi = \varphi_\alpha\varphi_\alpha\varphi_\beta$. In accordance with the possible distributions of the particles $1, 2, 3$ to the wave functions one gets the three functions

$$\psi_1 = \varphi_\alpha(\vec{\mathbf{r}}_1)\varphi_\alpha(\vec{\mathbf{r}}_2)\varphi_\beta(\vec{\mathbf{r}}_3)\,, \psi_2 = \varphi_\alpha(\vec{\mathbf{r}}_1)\varphi_\beta(\vec{\mathbf{r}}_2)\varphi_\alpha(\vec{\mathbf{r}}_3)\,, \psi_3 = \varphi_\beta(\vec{\mathbf{r}}_1)\varphi_\alpha(\vec{\mathbf{r}}_2)\varphi_\alpha(\vec{\mathbf{r}}_3)\,, \tag{5.106a}$$

which, when realizing permutations, transform among one another according to 5.3.3.1, **2.**, p. 337. This way one gets for the functions $\psi_1\psi_2\psi_3$ a three dimensional representation of the symmetric group S_3. According to (5.93) the matrix elements of the representation matrices can be found by investigating the action of the group elements (5.84) on the coordinate subscripts in the basis elements e_i. For example:

$$\begin{aligned} p_1\psi_1 &= p_1\varphi_\alpha(\vec{\mathbf{r}}_1)\varphi_\alpha(\vec{\mathbf{r}}_2)\varphi_\beta(\vec{\mathbf{r}}_3) = \varphi_\alpha(\vec{\mathbf{r}}_1)\varphi_\beta(\vec{\mathbf{r}}_2)\varphi_\alpha(\vec{\mathbf{r}}_3) = D_{21}(p_1)\psi_2,\\ p_1\psi_2 &= p_1\varphi_\alpha(\vec{\mathbf{r}}_1)\varphi_\beta(\vec{\mathbf{r}}_2)\varphi_\alpha(\vec{\mathbf{r}}_3) = \varphi_\alpha(\vec{\mathbf{r}}_1)\varphi_\alpha(\vec{\mathbf{r}}_2)\varphi_\beta(\vec{\mathbf{r}}_3) = D_{12}(p_1)\psi_1,\\ p_1\psi_3 &= p_1\varphi_\beta(\vec{\mathbf{r}}_1)\varphi_\alpha(\vec{\mathbf{r}}_2)\varphi_\alpha(\vec{\mathbf{r}}_3) = \varphi_\beta(\vec{\mathbf{r}}_1)\varphi_\alpha(\vec{\mathbf{r}}_2)\varphi_\alpha(\vec{\mathbf{r}}_3) = D_{33}(p_1)\psi_3. \end{aligned} \tag{5.106b}$$

Altogether one finds:

$$\begin{aligned} \mathbf{D}(e) &= \begin{pmatrix} 1&0&0\\0&1&0\\0&0&1 \end{pmatrix}, & \mathbf{D}(p_1) &= \begin{pmatrix} 0&1&0\\1&0&0\\0&0&1 \end{pmatrix}, & \mathbf{D}(p_2) &= \begin{pmatrix} 0&0&1\\0&1&0\\1&0&0 \end{pmatrix},\\ \mathbf{D}(p_3) &= \begin{pmatrix} 1&0&0\\0&0&1\\0&1&0 \end{pmatrix}, & \mathbf{D}(p_4) &= \begin{pmatrix} 0&1&0\\0&0&1\\1&0&0 \end{pmatrix}, & \mathbf{D}(p_5) &= \begin{pmatrix} 0&0&1\\1&0&0\\0&1&0 \end{pmatrix}. \end{aligned} \tag{5.106c}$$

For the characters one has:

$$\chi(e) = 3, \;\; \chi(p_1) = \chi(p_2) = \chi(p_3) = 1, \;\; \chi(p_4) = \chi(p_5) = 0. \tag{5.106d}$$

5.3.4.3 Direct Sum of Representations

The representations $D^{(1)}(G)$, $D^{(2)}(G)$ of dimension n_1 and n_2 can be composed to create a new representation $D(G)$ of dimension $n = n_1 + n_2$ by forming the direct sum of the representation matrices:

$$\mathbf{D}(a) = \mathbf{D}^{(1)}(a) \oplus \mathbf{D}^{(2)}(a) = \begin{pmatrix} \mathbf{D}^{(1)}(a) & \mathbf{0}\\ \mathbf{0} & \mathbf{D}^{(2)}(a) \end{pmatrix}. \tag{5.107}$$

The block-diagonal form of the representation matrix implies that the representation space V_n is the direct sum of two invariant subspaces $\mathrm{V}_{n_1}, \mathrm{V}_{n_2}$:

$$\mathrm{V}_n = \mathrm{V}_{n_1} \oplus \mathrm{V}_{n_2}, \quad n = n_1 + n_2. \tag{5.108}$$

A subspace V_m $(m < n)$ of V_n is called an invariant subspace if for any linear transformation $D(a)$, $a \in G$, every vector $\underline{\mathbf{x}} \in \mathrm{V}_m$ is mapped onto an element of V_m again:

$$\underline{\mathbf{x}}' = \mathbf{D}(a)\underline{\mathbf{x}} \quad \text{with} \quad \underline{\mathbf{x}}, \underline{\mathbf{x}}' \in \mathrm{V}_m. \tag{5.109}$$

The *character of the representation* (5.107) is the sum of the characters of the single representations:

$$\chi(a) = \chi^{(1)}(a) + \chi^{(2)}(a). \tag{5.110}$$

5.3.4.4 Direct Product of Representations

If $\underline{\mathbf{e}}_i$ $(i = 1, 2, \ldots, n_1)$ and $\underline{\mathbf{e}}'_k$ $(k = 1, 2, \ldots, n_2)$ are the basis vectors of the representation spaces V_{n_1} and V_{n_2}, respectively, then the tensor product

$$\underline{\mathbf{e}}_{ik} = \{\underline{\mathbf{e}}_i \underline{\mathbf{e}}_k\} \quad (i = 1, 2, \ldots, n_1;\ k = 1, 2, \ldots, n_2) \tag{5.111}$$

forms a basis in the product space $\mathrm{V}_{n_1} \otimes \mathrm{V}_{n_2}$ of dimension $n_1 \cdot n_2$. With the representations $D^{(1)}(G)$ and $D^{(2)}(G)$ in V_{n_1} and V_{n_2}, respectively an $n_1 \cdot n_2$-dimensional representation $D(G)$ in the product space can be constructed by forming the direct or (inner) Kronecker product (see 4.1.5,**9.**, p. 276) of the representation matrices:

$$\mathbf{D}(G) = \mathbf{D}^{(1)}(G) \otimes \mathbf{D}^{(2)}(G), \quad (D(G))_{ik,jl} = D^{(1)}_{ik}(a) \cdot D^{(2)}_{jl}(a)$$
$$\text{with} \quad i, k = 1, 2, \ldots, n_1;\ j, l = 1, 2, \ldots, n_2. \tag{5.112}$$

The character of the Kronecker product of two representations is equal to the product of the characters of the factors

$$\chi^{(1\times 2)}(a) = \chi^{(1)}(a) \cdot \chi^{(2)}(a). \tag{5.113}$$

5.3.4.5 Reducible and Irreducible Representations

If the representation space V_n possesses a subspace V_m $(m < n)$ invariant under the group operations the representation matrices can be decomposed according to

$$\mathbf{T}^{-1} \cdot \mathbf{D}(a) \cdot \mathbf{T} = \begin{pmatrix} \mathbf{D}_1(a) & \mathbf{A} \\ \mathbf{0} & \mathbf{D}_2(a) \end{pmatrix} \begin{matrix} \{\ m \text{ rows} \\ \{\ n-m \text{ rows} \end{matrix} \tag{5.114}$$

by a suitable transformation $\mathbf{T}$ of the basis in V_n. $\mathbf{D}_1(a)$ and $\mathbf{D}_2(a)$ themselves are matrix representations of $a \in G$ of dimension m and $n - m$, respectively.

A representation $D(G)$ is called *irreducible* if there is no proper (non-trivial) invariant subspace in V_n. The number of non-equivalent irreducible representations of a finite group is finite. If a transformation $\mathbf{T}$ of a basis can be found which makes V_n to a direct sum of invariant subspaces, i.e.,

$$\mathrm{V}_n = \mathrm{V}_1 \oplus \cdots \oplus \mathrm{V}_{n_j}, \tag{5.115}$$

then for every $a \in G$ the representation matrix $\mathbf{D}(a)$ can be transformed into the block-diagonal form ($\mathbf{A} = \mathbf{0}$ in (5.114)):

$$\mathbf{T}^{-1} \cdot \mathbf{D}(a) \cdot \mathbf{T} = \mathbf{D}^{(1)}(a) \oplus \cdots \oplus \mathbf{D}^{(n_j)}(a) = \begin{pmatrix} \mathbf{D}^{(1)}(a) & & 0 \\ & \ddots & \\ 0 & & \mathbf{D}^{(n_j)}(a) \end{pmatrix}. \tag{5.116}$$

by a similarity transformation with $\mathbf{T}$. Such a representation is called *completely reducible*.

Remark: For the application of group theory in natural sciences a fundamental task consists in the classification of all non-equivalent irreducible representations of a given group.

■ The representation of the symmetric group S_3 given in (5.106c), p. 343, is reducible. For example, in the basis transformation $\{\underline{\mathbf{e}}_1, \underline{\mathbf{e}}_2, \underline{\mathbf{e}}_3\} \longrightarrow \{\underline{\mathbf{e}}'_1 = \underline{\mathbf{e}}_1 + \underline{\mathbf{e}}_2 + \underline{\mathbf{e}}_3,\ \underline{\mathbf{e}}'_2 = \underline{\mathbf{e}}_2,\ \underline{\mathbf{e}}'_3 = \underline{\mathbf{e}}_3\}$ one obtains for the representation matrix of the permutation p_3 (with $\psi_1 = \underline{\mathbf{e}}_1, \psi_2 = \underline{\mathbf{e}}_2, \psi_3 = \underline{\mathbf{e}}_3$):

$$\mathbf{D}(p_3) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} = \begin{pmatrix} \mathbf{D_1}(p_3) & \mathbf{0} \\ \mathbf{A} & \mathbf{D_2}(p_3) \end{pmatrix} \tag{5.117}$$

with $\mathbf{A} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, $\mathbf{D_1}(p_3) = 1$ as the identity representation of S_3 and $\mathbf{D}_2(p_3) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$.

5.3.4.6 Schur's Lemma 1

If $\mathbf{C}$ is an operator commuting with all transformations of an irreducible representation $\mathbf{D}$ of a group $[\mathbf{C}, \mathbf{D}(a)] = \mathbf{C} \cdot \mathbf{D}(a) - \mathbf{D}(a) \cdot \mathbf{C} = 0, \ \ a \in G$, and the representation space V_n is an invariant subspace of $\mathbf{C}$, then $\mathbf{C}$ is a multiple of the unit operator, i.e., a matrix (c_{ik}) which commutates with all matrices of an irreducible representation is a multiple of the matrix $\mathbf{I}$, $\mathbf{C} = \lambda \cdot \mathbf{I}, \ \ \lambda \in \mathbb{C}$.

5.3.4.7 Clebsch-Gordan Series

In general, the Kronecker product of two irreducible representations $\mathbf{D}^{(1)}(G)$, $\mathbf{D}^{(2)}(G)$ is reducible. By a suitable basis transformation in the product space $\mathbf{D}^{(1)}(G) \otimes \mathbf{D}^{(2)}(G)$ can be decomposed into the direct sum of its irreducible parts $\mathbf{D}^{(\alpha)}$ $(\alpha = 1, 2, \ldots, n)$ (*Clebsch–Gordan theorem*). This expansion is called the *Clebsch–Gordan series*:

$$\mathbf{D}^{(1)}(G) \otimes \mathbf{D}^{(2)}(a) = \sum_{\alpha=1}^{n} \oplus \, m_\alpha \mathbf{D}^{(\alpha)}(G). \tag{5.118}$$

Here, m_α is the multiplicity with which the irreducible representation $\mathbf{D}^{(\alpha)}(G)$ occurs in the Clebsch–Gordan series.
The matrix elements of the basis transformation in the product space causing the reduction of the Kronecker product into its irreducible components are called *Clebsch–Gordan coefficients*.

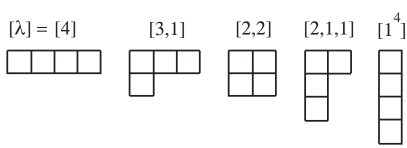
5.3.4.8 Irreducible Representations of the Symmetric Group S_M

1. Symmetric Group S_M
The non-equivalent irreducible representations of the symmetric group S_M are characterized uniquely by the partitions of M, i.e., by the splitting of M into integers according to

$$[\lambda] = [\lambda_1, \lambda_2, \ldots, \lambda_M], \quad \lambda_1 + \lambda_2 + \cdots + \lambda_M = M, \quad \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_M \geq 0. \tag{5.119}$$

The graphic representation of the partitions is done by arranging boxes in *Young diagrams*.

■ For the group S_4 one obtains five Young diagrams as shown in the figure.



The dimension of the representation $[\lambda]$ is given by

$$n^{[\lambda]} = M! \frac{\prod\limits_{i<j\leq k} (\lambda_i - \lambda_j + j - i)}{\prod_{i=1}^{k} (\lambda_i + k - i)!}. \tag{5.120}$$

The Young diagram $[\tilde{\lambda}]$ conjugated to $[\lambda]$ is constructed by the interchange of rows and columns. In general, the irreducible representation of S_M is reducible if one restricts to one of the subgroups $S_{M-1}, S_{M-2}, \cdots$.

■ In quantum mechanics for a system of identical particles the Pauli principle demands the construction of many-body wave functions that are antisymmetric with respect to the interchange of all coordinates of two arbitrary particles. Often, the wave function is given as the product of a function in space coordinates and a function in spin variables. If for such a case due to particle permutations the spatial part of the wave function transforms according to the irreducible representation $[\lambda]$ of the symmetric group, then it has to be combined with a spin function transforming according to $[\tilde{\lambda}]$ in order to get a total wave function which is antisymmetric if two particles are interchanged.

5.3.5 Applications of Groups

In chemistry and in physics, groups are applied to describe the "symmetry" of the corresponding objects. Such objects are, for instance, molecules, crystals, solid structures or quantum mechanical

systems. The basic idea of these applications is the von Neumann principle:
If a system has a certain group of symmetry operations, then every physical observational quantity of this system must have the same symmetry.

5.3.5.1 Symmetry Operations, Symmetry Elements

A *symmetry operation* s of a space object is a mapping of the space into itself such that the length of line segments remains unchanged and the object goes into a covering position to itself. The set of fixed points of the symmetry operation s is denoted by Fix s, i.e., the set of all points of space which remain unchanged for s. The set Fix s is called the *symmetry element* of s. The Schoenflies symbolism is used to denote the symmetry operation.
Two types of symmetry operations are distinguished: Operations without a fixed point and operations with at least one fixed point.

1. Symmetry Operations without a Fixed Point, for which no point of the space stays unchanged, cannot occur for bounded space objects, but now only such objects are considered. A symmetry operation without a fixed point is for instance a parallel translation.

2. Symmetry Operations with at least One Fixed Point are for instance rotations and reflections. The following operations belong to them.

a) Rotations Around an Axis by an Angle φ: The axis of rotation and also the rotation itself is denoted by C_n for $\varphi = 2\pi/n$. The axis of rotation is then called of n-th order.

b) Reflection with Respect to a Plane: Both the plane of reflection and the reflection itself are denoted by σ. If additionally there is a principal rotation axis, then one draws it perpendicularly and denote the planes of reflections which are perpendicular to this axis by σ_h (h from horizontal) and the planes of reflections passing through the rotational axis are denoted by σ_v (v from vertical) or σ_d (d means dihedral, if certain angles are halved).

c) Improper Orthogonal Mappings: An operation such that after a rotation C_n a reflection σ_h follows, is called an improper orthogonal mapping and it is denoted by S_n. Rotation and reflection commute. The axis of rotation is then called an improper rotational axis of n-th order and it is also denoted by S_n. This axis is called the corresponding symmetry element, although only the symmetry center stays fixed under the application of the operation S_n. For $n = 2$, an improper orthogonal mapping is also called a point reflection or inversion (see 4.3.5.1, p. 287) and it is denoted by i.

5.3.5.2 Symmetry Groups or Point Groups

For every symmetry operation S, there is an inverse operation S^{-1}, which reverses S "back", i.e.,

$$SS^{-1} = S^{-1}S = \epsilon. \tag{5.121}$$

Here ϵ denotes the identity operation, which leaves the whole space unchanged. The family of symmetry operations of a space object forms a group with respect to the successive application, which is in general a non-commutative *symmetry group* of the objects. The following relations hold:

a) Every rotation is the product of two reflections. The intersection line of the two reflection planes is the rotation axis.

b) For two reflections σ and σ'

$$\sigma\sigma' = \sigma'\sigma \tag{5.122}$$

if and only if the corresponding reflection planes are identical or they are perpendicular to each other. In the first case the product is the identity ϵ, in the second one the rotation C_2.

c) The product of two rotations with intersecting rotational axes is again a rotation whose axis goes through the intersection point of the given rotational axes.

d) For two rotations C_2 and C_2' around the same axis or around axes perpendicular to each other:

$$C_2C_2' = C_2'C_2. \tag{5.123}$$

The product is again a rotation. In the first case the corresponding rotational axis is the given one, in the second one the rotational axis is perpendicular to the given ones.

5.3.5.3 Symmetry Operations with Molecules

It requires a lot of work to recognize every symmetry element of an object. In the literature, for instance in [5.10], [5.13], it is discussed in detail how to find the symmetry groups of molecules if all the symmetry elements are known. The following notation is used for the interpretation of a molecule in space: The symbols above C in **Fig. 5.11** mean that the OH group lies above the plane of the drawing, the symbol to the right-hand side of C means that the group OC_2H_5 is under C.
The determination of the symmetry group can be made by the following method.

1. No Rotational Axis

a) If no symmetry element exists, then $G = \{\epsilon\}$ holds, i.e., the molecule does not have any symmetry operation but the identity ϵ.

■ The molecule hemiacetal **(Fig.5.11)** is not planar and it has four different atom groups.

b) If σ is a reflection or i is an inversion, then $G = \{\epsilon, \sigma\} =: C_s$ or $G = \{\epsilon, i\} = C_i$ hold, and with this it is isomorphic to Z_2.

■ The molecule of tartaric acid **(Fig.5.12)** can be reflected in the center P (inversion).

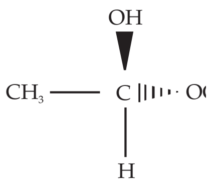


Figure 5.11

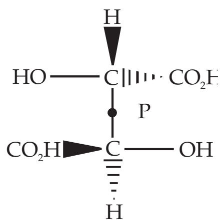


Figure 5.12

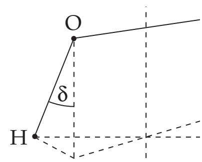


Figure 5.13

2. There is Exactly One Rotational Axis *C*

a) If the rotation can have any angle, i.e., $C = C_\infty$, then the molecule is linear, and the symmetry group is infinite.

■ **A:** For the molecule of sodium chloride (common salt) NaCl there is no horizontal reflection. The corresponding symmetry group of all the rotations around C is denoted by $C_{\infty v}$.

■ **B:** The molecule O_2 has one horizontal reflection. The corresponding symmetry group is generated by the rotations and by this reflection, and it is denoted by $D_{\infty h}$.

b) The rotation axis is of n-th order, $C = C_n$, but it is not an improper rotational axis of order $2n$.
If there is no further symmetry element, then G is generated by a rotation d by an angle π/n around C_n, i.e., $G = < d > \cong Z_n$. In this case G is also denoted by C_n.
If there is a further vertical reflection σ_v, then $G = < d, \sigma_v > \cong D_n$ holds (see 5.3.3.1, p. 336), and G is denoted by C_{nv}.
If there exists an additional horizontal reflection σ_h, then $G = < d, \sigma_v > \cong Z_n \times Z_2$ holds. G is denoted by C_{nh} and it is cyclic for odd n (see 5.3.3.2, p. 337).

■ **A:** For hydrogen peroxide **(Fig.5.13)** these three cases occur in the order given above for $0 < \delta < \pi/2, \delta = 0$ and $\delta = \pi/2$.

■ **B:** The molecule of water H_2O has a rotational axis of second order and a vertical plane of reflection, as symmetry elements. Consequently, the symmetry group of water is isomorphic to the group D_2, which is isomorphic to the Klein four-group V_4 (see 5.3.3.2, **3.**, p. 338).

c) The rotational axis is of order n and at the same time it is also an improper rotational axis of order

$2n$. We have to distinguish two cases.
$\boldsymbol{\alpha}$) There is no further vertical reflection, so $G \cong Z_{2n}$ holds, and G is denoted also by S_{2n}.
■ An example is the molecule of tetrahydroxy allene with formula $C_3(OH)_4$ **(Fig.5.14)**.
$\boldsymbol{\beta}$) If there is a vertical reflection, then G is a group of order $4n$, which is denoted by D_{2n}.
■ $n = 2$ gives $G \cong D_4$, i.e., the dihedral group of order eight. An example is the allene molecule **(Fig.5.15)**.

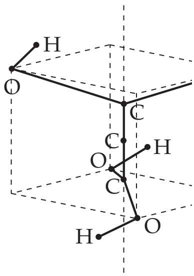


Figure 5.14

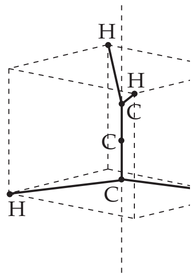


Figure 5.15

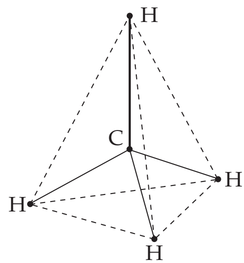


Figure 5.16

3. Several Rotational Axes If there are several rotational axes, then one has to distinguish further cases. In particular, if several rotational axes have an order $n \geq 3$, then the following groups are the corresponding symmetry groups.
a) Tetrahedral group $\boldsymbol{T_d}$: Isomorphic to S_4, $\mathrm{ord}T_d = 24$.
b) Octahedral group $\boldsymbol{O_h}$: Isomorphic to $S_4 \times Z_2$, $\mathrm{ord}O_h = 48$.
c) Icosahedral group $\boldsymbol{I_h}$: $\mathrm{ord}I_h = 120$.
These groups are the symmetry groups of the regular polyhedron discussed in 3.3.3, **Table 3.7**, p. 155, **(Fig.3.63)**.
■ The methane molecule **(Fig.5.16)** has the tetrahedral group T_d as a symmetry group.

5.3.5.4 Symmetry Groups in Crystallography

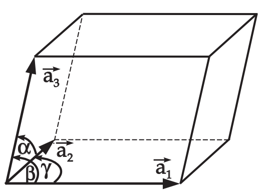


Figure 5.17

1. Lattice Structures
In crystallography the parallelepiped represents, independently of the arrangement of specific atoms or ions, the elementary (unit) cell of the *crystal lattice*. It is determined by three non-coplanar basis vectors $\vec{\mathbf{a}}_i$ starting from one lattice point **(Fig. 5.17)**. The infinite geometric lattice structure is created by performing all *primitive translations* $\vec{\mathbf{t}}_n$:

$$\vec{\mathbf{t}}_{\mathbf{n}} = n_1\vec{\mathbf{a}}_1 + n_2\vec{\mathbf{a}}_2 + n_3\vec{\mathbf{a}}_3, \quad n = (n_1, n_2, n_3) \quad n_i \in \mathrm{Z}. \tag{5.124}$$

Here, the coefficients n_i $(i = 1, 2, \ldots)$ are integers.

All the translations $\vec{\mathbf{t}}_n$ fixing the space points of the lattice $L = \{\vec{\mathbf{t}}_n\}$ in terms of lattice vectors form the translation group T with the group element $T(\vec{\mathbf{t}}_n)$, the inverse element $T^{-1}(\vec{\mathbf{t}}_{\mathbf{n}}) = T(-\vec{\mathbf{t}}_{\mathbf{n}})$, and the composition law $T(\vec{\mathbf{t}}_{\mathbf{n}}) * T(\vec{\mathbf{t}}_{\mathbf{m}}) = T(\vec{\mathbf{t}}_{\mathbf{n}} + \vec{\mathbf{t}}_{\mathbf{m}})$. The application of the group element $T(\vec{\mathbf{t}}_{\mathbf{n}})$ to the position vector $\vec{\mathbf{r}}$ is described by:

$$T(\vec{\mathbf{t}}_{\mathbf{n}})\vec{\mathbf{r}} = \vec{\mathbf{r}} + \vec{\mathbf{t}}_{\mathbf{n}}. \tag{5.125}$$

2. Bravais Lattices

Taking into account the possible combinations of the relative lengths of the basis vectors $\vec{\mathbf{a_i}}$ and the pairwise related angles between them (particularly angles 90° and 120°) one obtains seven different types of *elementary cells* with the corresponding lattices, the *Bravais lattices* (see **Fig. 5.17**, and **Table 5.4**). This classification can be extended by seven *non-primitive elementary cells* and their corresponding lattices by adding additional lattice points at the intersection points of the face or body diagonals, preserving the symmetry of the elementary cell. In this way one may distinguish one-side face-centered lattices, body-centered lattices, and all-face centered lattices.

Tabelle 5.4 Primitive Bravais lattice

Elementary cell	Relative lengths of basis vectors	Angles between basis vectors
triclinic	$a_1 \neq a_2 \neq a_3$	$\alpha \neq \beta \neq \gamma \neq 90°$
monoclinic	$a_1 \neq a_2 \neq a_3$	$\alpha = \gamma = 90° \neq \beta$
rhombic	$a_1 \neq a_2 \neq a_3$	$\alpha = \beta = \gamma = 90°$
trigonal	$a_1 = a_2 = a_3$	$\alpha = \beta = \gamma < 120° (\neq 90°)$
hexagonal	$a_1 = a_2 \neq a_3$	$\alpha = \beta = 90°, \gamma = 120°$
tetragonal	$a_1 = a_2 \neq a_3$	$\alpha = \beta = \gamma = 90°$
cubic	$a_1 = a_2 = a_3$	$\alpha = \beta = \gamma = 90°$

3. Symmetry Operations in Crystal Lattice Structures

Among the symmetry operations transforming the space lattice to equivalent positions there are point group operations such as certain rotations, improper rotations, and reflections in planes or points. But not all point groups are also crystallographic point groups. The requirement that the application of a group element to a lattice vector $\vec{\mathbf{t_n}}$ leads to a lattice vector $\vec{\mathbf{t'_n}} \in L$ (L is the set of all lattice points) again restricts the allowed point groups P with the group elements $P(R)$ according to:

$$P = \{R : R\vec{\mathbf{t_n}} \in L\}, \quad \vec{\mathbf{t_n}} \in L. \tag{5.126}$$

Here, R denotes a proper ($R \in SO(3)$) or improper rotation operator ($R = IR' \in O(3), R' \in SO(3), I$ is the inversion operator with $I\vec{\mathbf{r}} = -\vec{\mathbf{r}}$, $\vec{\mathbf{r}}$ is a position vector). For example, only n-fold rotation axes with $n = 1, 2, 3, 4$ or 6 are compatible with a lattice structure. Altogether, there are 32 crystallographic point groups P.

The symmetry group of a space lattice may also contain operators representing simultaneous applications of rotations and primitive translations. In this way one gets gliding reflections, i.e., reflections in a plane and translations parallel to the plane, and screws, i.e., rotations through $2\pi/n$ and translations by $m\vec{\mathbf{a}}/n$ ($m = 1, 2, \ldots, n-1$, $\vec{\mathbf{a}}$ are basis translations). Such operations are called non-primitive translations $\vec{\mathbf{V}}(R)$, because they correspond to "fractional" translations. For a gliding reflection R is a reflection and for a screw R is a proper rotation.

The elements of the space group G, for which the crystal lattice is invariant are composed of elements P of the crystallographic point group P, primitive translations $T(\vec{\mathbf{t_n}})$ and non-primitive translations $\vec{\mathbf{V}}(R)$:

$$G = \{\{R|\vec{\mathbf{V}}(R) + \vec{\mathbf{t_n}} : R \in P, \quad \vec{\mathbf{t_n}} \in L\}\}. \tag{5.127}$$

The unit element of the space group is $\{e|0\}$ where e is the unit element of R. The element $\{e|\vec{\mathbf{t_n}}\}$ means a primitive translation, $\{R|0\}$ represents a rotation or reflection. Applying the group element

$\{R|\vec{\mathbf{t}}_n\}$ to the position vector $\vec{\mathbf{r}}$ one obtains:

$$\{R|\vec{\mathbf{t_n}}\}\vec{\mathbf{r}} = R\vec{\mathbf{r}} + \vec{\mathbf{t_n}}. \tag{5.128}$$

4. Crystal Systems (Holohedry)

From the 14 Bravais lattices, $L = \{\vec{\mathbf{t_n}}\}$, the 32 crystallographic point groups $P = \{R\}$ and the allowed non-primitive translations $\vec{\mathbf{V}}(R)$ one can construct 230 space groups $G = \{R|\vec{\mathbf{V}}(R) + \vec{\mathbf{t_n}}\}$. The point groups correspond to 32 crystallographic classes. Among the point groups there are seven groups that are not a subgroup of another point group but contain further point groups as a subgroup. Each of these seven point groups form a *crystal system* (*holohedry*). The symmetry of the seven crystal systems is reflected in the symmetry of the seven Bravais lattices. The relation of the 32 crystallographic classes to the seven crystal systems is given in **Table 5.5** using the notation of Schoenflies.

Remark: The space group G (5.127) is the symmetry group of the "empty" lattice. The real crystal is obtained by arranging certain atoms or ions at the lattice sites. The arrangement of these crystal constituents exhibits its own symmetry. Therefore, the symmetry group G_0 of the real crystal possesses a lower symmetry than G $(G \supset G_0)$, in general.

Table 5.5 Bravais lattice, crystal systems, and crystallographic classes
Notation: C_n – rotation about an n-fold rotation axis, D_n – dihedral group, T_n – tetrahedral group, O_n – octahedral group, S_n – mirror rotations with an n-fold axis.

Lattice type	**Crystal system (holohedry)**	**Crystallographic class**
triclinic	C_i	C_1, C_i
monoclinic	C_{2h}	C_2, C_h, C_{2h}
rhombic	D_{2h}	C_{2v}, D_2, D_{2h}
tetragonal	D_{4h}	$C_4, S_4, C_{4h}, D_4, C_{4v}, D_{2d}, D_{4h}$
hexagonal	D_{6h}	$C_6, C_{3h}, C_{6h}, D_6, C_{6v}, D_{3h}, D_{6h}$
trigonal	D_{3d}	$C_3, S_6, D_3, C_{3v}, D_{3d}$
cubic	O_h	T, T_h, T_d, O, O_h

5.3.5.5 Symmetry Groups in Quantum Mechanics

Linear coordinate transformations that leave the Hamiltonian $\hat{H}$ of a quantum mechanical system (see 9.2.4, **2.**, p. 593) invariant represent a symmetry group G, whose elements g commute with $\hat{H}$:

$$[g, \hat{H}] = g\hat{H} - \hat{H}g = 0, \quad g \in G. \tag{5.129}$$

The commutation property of g and $\hat{H}$ implies that in the application of the product of the operators g and $\hat{H}$ to a state φ the sequence of the action of the operators is arbitrary:

$$g(\hat{H}\varphi) = \hat{H}(g\varphi). \tag{5.130}$$

Hence, one has: If $\varphi_{E\alpha}$ $(\alpha = 1, 2, \ldots, n)$ are the eigenstates of $\hat{H}$ with energy eigenvalue E of degeneracy n, i.e.,

$$\hat{H}\varphi_{E\alpha} = E\varphi_{E\alpha} \quad (\alpha = 1, 2, \ldots, n), \tag{5.131}$$

then the transformed states $g\varphi_{E\alpha}$ are also eigenstates belonging to the same eigenvalue E:

$$g\hat{H}\varphi_{E\alpha} = \hat{H}g\varphi_{E\alpha} = Eg\varphi_{E\alpha}. \tag{5.132}$$

The transformed states $g\varphi_{E\alpha}$ can be written as a linear combination of the eigenstates $\varphi_{E\alpha}$:

$$g\varphi_{E\alpha} = \sum_{\beta=1}^{n} D_{\beta\alpha}(g)\varphi_{E\beta}. \tag{5.133}$$

Hence, the eigenstates $\varphi_{E\alpha}$ form the basis of an n-dimensional representation space for the representation $D(G)$ of the symmetry group G of the Hamiltonian $\hat{H}$ with the representation matrices $(D_{\alpha\beta}(g))$. This representation is irreducible if there are no "hidden" symmetries. One can state that the energy eigenstates of a quantum mechanical system can be labeled by the signatures of the irreducible representations of the symmetry group of the Hamiltonian.

Thus, the representation theory of groups allows for qualitative statements on such patterns of the energy spectrum of a quantum mechanical system which are established by the outer or inner symmetries of the system only. Also the splitting of degenerate energy levels under the influence of a perturbation which breaks the symmetry or the selection rules for the matrix elements of transitions between energy eigenstates follows from the investigation of representations according to which the participating states and operators transform under group operations.

The application of group theory in quantum mechanics is presented extensively in the literature (see, e.g., [5.6], [5.7], [5.8], [5.10], [5.11]).

5.3.5.6 Further Applications of Group Theory in Physics

Further examples of the application of particular continuous groups in physics can only be mentioned here (see, e.g., [5.6], [5.10]).

$U(1)$: Gauge transformations in electrodynamics.

$SU(2)$: Spin and isospin multiplets in particle physics.

$SU(3)$: Classification of the baryons and mesons in particle physics. Many-body problem in nuclear physics.

$SO(3)$: Angular momentum algebra in quantum mechanics. Atomic and nuclear many-body problems.

$SO(4)$: Degeneracy of the hydrogen spectrum.

$SU(4)$: Wigner super-multiplets in the nuclear shell model due to the unification of spin and isospin degrees of freedom. Description of flavor multiplets in the quark model including the charm degree of freedom.

$SU(6)$: Multiplets in the quark model due to the combination of flavor and spin degrees of freedom. Nuclear structure models.

$U(n)$: Shell models in atomic and nuclear physics.

$SU(n), SO(n)$: Many-body problems in nuclear physics.

$SU(2) \otimes U(1)$: Standard model of the electro weak interaction.

$SU(5) \supset SU(3) \otimes SU(2) \otimes U(1)$: Unification of fundamental interactions (GUT).

Remark: The groups $SU(n)$ and $SO(n)$ are Lie groups, i.e. continuous groups (see, 5.3.6, p. 351 and e.g., [5.6]).

5.3.6 Lie Groups and Lie Algebras

5.3.6.1 Introduction

Lie groups and *Lie algebras* are named after the Norwegian mathematician Sophus Lie (1842-1899). In this chapter only Lie groups of matrices are considered since they are most important in applications. Main examples of matrix-Lie groups are:

- the group $O(n)$ of orthogonal matrices,
- the subgroup $SO(n)$ of orthogonal matrices of determinants $+1$, i.e. the orthogonal matrices describing rotations in $\mathbb{R}^n$,

- the Euclidean group $SE(n)$, which describes rigid-body motions.

These groups have many applications in computer graphics and in robotics.

The most important relation between a Lie group and the corresponding Lie algebra will be described by the exponential mapping. This relation is explained by the following example.

■ The solution of initial value problems of first order differential equations or of a system of differential equations can be determined with the help of the exponential function.
The initial value problem (5.134a) for $y = y(t)$ has the following solution (5.134b):

$$\frac{dy}{dt} = x\,y \quad (x \text{ const}) \text{ with } y(0) = y_0\,, \qquad (5.134a) \qquad\qquad y(t) = e^{xt} y_0\,. \qquad (5.134b)$$

Similarly, for the system of first order differential equations with unknown vector $\vec{\mathbf{y}} = \vec{\mathbf{y}}(t)$ and with the constant coefficient matrix $\mathbf{X}$ the initial value problem (5.135a)

$$\frac{d\vec{\mathbf{y}}}{dt} = \left(\frac{dy_1}{dt}, \frac{dy_2}{dt}, \ldots, \frac{dy_n}{dt}\right)^{\mathrm{T}} = \mathbf{X}\vec{\mathbf{y}} \quad (\text{matrix } \mathbf{X} \text{ const}) \text{ with } \vec{\mathbf{y}}(0) = \vec{\mathbf{y}}_0, \tag{5.135a}$$

has the solution (5.135b) with the matrix-exponential function $e^{t\mathbf{X}}$:

$$\vec{\mathbf{y}}(t) = e^{\mathbf{X}\,t}\vec{\mathbf{y}}_0\,, \qquad e^{t\mathbf{X}} := \sum_{k=0}^{\infty} \frac{1}{k!} t^k \mathbf{X}^k = I_{n\times n} + \sum_{k=1}^{\infty} \frac{1}{k!} t^k \mathbf{X}^k\,. \tag{5.135b}$$

The special matrix-exponential function $e^{t\mathbf{X}}$ for a given quadratic $n \times n$ matrix $\mathbf{X}$ has the following properties:

- $e^{0\mathbf{X}} = I_{n\times n}$, where $I_{n\times n}$ denotes the unit matrix.
- $e^{t\mathbf{X}}$ is invertible, because $\det e^{t\mathbf{X}} = e^{t\cdot \mathrm{Spur}\,\mathbf{X}} \neq 0$.
- $e^{t_1\mathbf{X}} e^{t_2\mathbf{X}} = e^{(t_1+t_2)\mathbf{X}} = e^{t_2\mathbf{X}} e^{t_1\mathbf{X}}$ for every $t_1,\ t_2 \in \mathbb{R}$, but in general is $e^{\mathbf{X}_1} e^{\mathbf{X}_2} \neq e^{\mathbf{X}_2} e^{\mathbf{X}_1} \neq e^{\mathbf{X}_1+\mathbf{X}_2}$.
- In particular $e^{-t\mathbf{X}} e^{t\mathbf{X}} = e^{t\mathbf{X}} e^{-t\mathbf{X}} = I_{n\times n}$.
- $\left.\frac{d}{dt} e^{t\mathbf{X}}\right|_{t=0} = \mathbf{X}\, e^{t\mathbf{X}}\big|_{t=0} = \mathbf{X}$.

Consequently, the elements $e^{t\mathbf{X}}$ (for a fixed $\mathbf{X}$) form a multiplicative group with respect to matrix multiplication. Since $t \in \mathbb{R}$, the matrices $e^{t\mathbf{X}}$ form a one dimensional group. At the same time it is one of the simplest examples of Lie groups. It will be shown that matrices $\mathbf{X}$ and $t\mathbf{X}$ are elements of the Lie algebra belonging to this Lie group (see 5.3.6.4, p. 356). In this way the exponential function generates the Lie group from the elements of the Lie algebra.

5.3.6.2 Matrix-Lie Groups

For matrix-Lie groups it is not necessary to define Lie groups in general. For general Lie groups there should be introduced the notion of differentiable manifolds, which is not needed here. For matrix-Lie groups the following definitions are important, while in further discussions the main topic will be the *general linear group*.

1. General Linear Group

1. Group A group (see 5.3.3, p. 336) is a set G with a map

$$G \times G \to G\,, \quad (g, h) \mapsto g * h\,, \tag{5.136a}$$

which is the so called group operation or group multiplication with the following properties:

- Associativity: for every $g, h, k \in G$

$$g * (h * k) = (g * h) * k\,, \tag{5.136b}$$

- Existence of identity: There is an element $e \in G$, such that for every $g \in G$

$$g * e = e * g = g\,, \tag{5.136c}$$

• Existence of an inverse: For every $g \in G$ there is an element $h \in G$ such that

$$g * h = h * g = e\,. \tag{5.136d}$$

Remark 1: If $g * h = h * g$ for every$g, h \in G$, then the group is called *commutative*. The matrix groups considered here are not commutative. It follows obviously from the definition, that the product of two elements of the group also belongs to the group, so the group is closed with respect to group multiplication.

Remark 2: Let $M_n(\mathbb{R})$ the vector space of all $n \times n$ matrices with real entries. $M_n(\mathbb{R})$ is obviously not a group with respect to matrix multiplication, since not every $n \times n$ matrix is invertible.

2. Definition of the General Linear Group The set of all real, invertible, $n \times n$ matrices, which obviously form a group with respect to matrix multiplication, is called the *general linear group* and is denoted by $GL(n, \mathbb{R})$.

2. Matrix-Lie Groups

1. Convergence of Matrices A sequence $\{\mathbf{A}_m\}_{m=1}^{\infty}$ of matrices $\mathbf{A}_m = (a_{kl}^{(m)})_{k,l=1}^{n}$ where $\mathbf{A}_m \in M_n(\mathbb{R})$ converges to the $n \times n$ matrix $\mathbf{A}$, if every sequence of entries $\{(a_{kl}^{(m)})\}_{m=1}^{\infty}$ converges to the corresponding matrix entry a_{kl} in the sense of convergence of real numbers.

2. Definition of the Matrix-Lie Groups A matrix-Lie group is a subgroup G of $GL(n, \mathbb{R})$ with the property: Let $\{\mathbf{A}_m\}_{m=1}^{\infty}$ be an arbitrary sequence of matrices from G converging to a matrix $\mathbf{A} \in M_n(\mathbb{R})$ in the sense of convergence in $M_n(\mathbb{R})$. Then either $\mathbf{A} \in G$ or $\mathbf{A}$ is not invertible.

This definition can be also formulated in the following way: A matrix-Lie group is a subgroup which is also a closed subset of $GL(n\,, \mathbb{R})$. (It does not mean, that G must be closed in $M_n(\mathbb{R})$).

3. Dimension of the Matrix-Lie Group The dimension of a matrix-Lie group is defined as the dimension of the corresponding Lie algebra (see 5.3.6.4, p. 356). The matrix-Lie group $GL(n\,, \mathbb{R})$ has dimension n^2.

3. Continuous Groups

Matrix-Lie groups can be introduced also with the help of continuous groups (see [22.22], [5.9], [5.7]).

1. Definition A continuous group is a special infinite group whose elements are given uniquely by a continuous parameter vector $\underline{\varphi} = (\varphi_1, \varphi_2, \ldots, \varphi_n)$:

$$a = a(\underline{\varphi})\,. \tag{5.137}$$

■ Group of rotation matrices in $\mathbb{R}^2$ (see (3.432), p. 230):

$$D = \begin{pmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{pmatrix} = a(\varphi) \ \text{ mit } \ 0 \le \varphi \le 2\pi\,. \tag{5.138}$$

The group elements depend only on one real parameter φ.

2. Product The product of two elements $a_1 = a(\underline{\varphi}_1)$, $a_2 = a(\underline{\varphi}_2)$ of a continuous group with elements $a = a(\underline{\varphi})$ is given by

$$a_1 * a_2 = a_3 = a(\underline{\boldsymbol{\varphi_3}}) \ \text{ with} \tag{5.139a}$$

$$\varphi_3 \ = \underline{f}(\underline{\boldsymbol{\varphi_1}}, \underline{\boldsymbol{\varphi_2}})\,, \tag{5.139b}$$

where the components of $\underline{f}(\underline{\boldsymbol{\varphi_1}}, \underline{\boldsymbol{\varphi_2}})$ are continuously differentiable functions.

■ The product of two rotation matrices $a = a(\varphi_1)$ and $a = a(\varphi_2)$ with $0 \le \varphi_1, \varphi_2 \le 2\pi$ ($a(\varphi)$ as in (5.138), is $a_3 = a(\varphi_1) * a(\varphi_2) = a(\varphi_3)$ with $\varphi_3 = f(\varphi_1, \varphi_2) = \varphi_1 + \varphi_2$. Using the Falk's scheme (see 4.1.4, **5.**, p. 273) and addition theorems one gets:

	$a(\varphi_2)$
$a(\varphi_1)$	$a(\varphi_3) = a(\varphi_1 + \varphi_2)$

or detailed

		$\cos\varphi_2$	$-\sin\varphi_2$
		$\sin\varphi_2$	$\cos\varphi_2$
$\cos\varphi_1$	$-\sin\varphi_1$	$\cos\varphi_1\cos\varphi_2 - \sin\varphi_1\sin\varphi_2$	$-\cos\varphi_1\sin\varphi_2 - \sin\varphi_1\cos\varphi_2$
$\sin\varphi_1$	$\cos\varphi_1$	$\sin\varphi_1\cos\varphi_2 + \cos\varphi_1\sin\varphi_2$	$-\sin\varphi_1\sin\varphi_2 + \cos\varphi_1\cos\varphi_2$

3. Dimension The parameter vectors $\underset{\sim}{\varphi}$ are elements of a vector space which is called parameter space. In this parameter space there is a domain which is given as the domain of the continuous group, and it is called the group space. The dimension of this group space is considered as the dimension of the continuous group.

■ **A:** The group of the real quadratic $n \times n$ invertible matrices has the dimension n^2, since every entry can be considered as a parameter.

■ **B:** The group of the rotation matrices (with respect to matrix multiplication) D in (5.138) has dimension 1. The rotation matrices are of type 2×2, but their four entries depend only on one parameter φ $(0 \le \varphi \le 2\pi)$.

4. Lie Groups

1. Definition of the Lie Group A Lie group is a continuous group where all elements of the group are given as continuous functions of the parameters.

2. Special Matrix-Lie Groups and their Dimension

■ **A Group $SO(n)$ of Rotations R:** The group $SO(n)$ of rotations $\mathbf{R}$ acts on the elements $\vec{\mathbf{x}} \in \mathbb{R}^n$ with matrix multiplication as $\vec{\mathbf{x}}\,' = \mathbf{R}\,\vec{\mathbf{x}} \in \mathbb{R}^n$. $SO(n)$ is an $n(n-1)/2$-dimensional Lie group.

■ **B Special Euclidean Group $SE(n)$:** The special Euclidian group $SE(n)$ consists of elements $g = (\mathbf{R}, \vec{\mathbf{b}})$ with $\mathbf{R} \in SO(n)$ and $\vec{\mathbf{b}} \in \mathbb{R}^n$ and with group multiplication $g_1 \circ g_2 = (\mathbf{R}_1\,\mathbf{R}_2,\ \mathbf{R}_1\vec{\mathbf{b}}_2 + \vec{\mathbf{b}}_1)$. It acts on the elements of Euclidean spaces $\mathbb{R}^n$ as

$$\vec{\mathbf{x}}\,' = \mathbf{R}\vec{\mathbf{x}} + \vec{\mathbf{b}}\,. \tag{5.140}$$

$SE(n)$ is the group of rigid-body motions of n-dimensional Euclidean space, it is an $n(n+1)/2$-dimensional Lie group. Discrete subgroups of $SE(n)$ are e.g. the crystallographic space groups, i.e. the symmetry group of a regular crystal-lattice.

■ **C Scaled Euclidean Group $SIM(n)$:** The scaled Euclidian group $SIM(n)$ consists of all pairs $(e^a\mathbf{R},\ \vec{\mathbf{b}})$ with $a \in \mathbb{R}$, $\mathbf{R} \in SO(n)$, $\vec{\mathbf{b}} \in \mathbb{R}^n$, with group multiplication $g_1 \circ g_2 = (e^{a_1+a_2}\mathbf{R}_1\,\mathbf{R}_2,\ \mathbf{R}_1\vec{\mathbf{b}}_2 + \vec{\mathbf{b}}_1)$. It acts on the elements of $\mathbb{R}^n$ by translation, rotation and dilatation (=stretching or shrinking):

$$\vec{\mathbf{x}}\,' = e^a\mathbf{R}\vec{\mathbf{x}} + \vec{\mathbf{b}}\,. \tag{5.141}$$

The scaled Euclidean group has the dimension $1 + n(n+1)/2$.

■ **D Real Special Linear Group $SL(n\,,\mathbb{R})$:** The real special linear group consists of all (real) $n\times n$ matrices with determinant +1. It acts on the elements of $\mathbb{R}^n$ with $\vec{\mathbf{x}}\,' = \mathbf{L}\vec{\mathbf{x}}$ by rotation, distortion and shearing so that the volume remains the same and parallel lines remain parallel. The dimension is $n^2 - 1$.

■ **E Special Affine Group:** The special affine groups of $\mathbb{R}^n$, which consists of all pairs $(e^a\,\mathbf{L},\ \vec{\mathbf{b}})$ with $L \in SL(n)$ and $\vec{\mathbf{b}} \in \mathbb{R}^n$, acts on the objects in $\mathbb{R}^n$ as rotation, translation, shearing, distortion and dilatation. This Lie group is the most general group of deformations in Euclidean spaces mapping parallel lines into parallel lines; it has dimension $n(n+1)$.

■ **F Group $SO(2)$:** The group $SO(2)$ describes all rotations about the origin in $\mathbb{R}^2$:

$$\mathrm{SO}(2) = \left\{ \begin{pmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{pmatrix},\ \varphi \in \mathbb{R} \right\} \tag{5.142}$$

■ **G Group $SL(2)$:** Every element of $SL(2)$ can be represented as

$$\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} e^t & 0 \\ 0 & e^{-t} \end{pmatrix} \begin{pmatrix} 1 & \xi \\ 0 & 1 \end{pmatrix} . \tag{5.143}$$

■ **H Group $SE(2)$:** The elements of the group $SE(2)$ can be represented as 3×3 matrices:

$$\begin{pmatrix} \cos\theta & -\sin\theta & x_1 \\ \sin\theta & \cos\theta & x_2 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{with } \theta \in \mathbb{R} \text{ and } \vec{\mathbf{x}} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^2 . \tag{5.144}$$

Remark: Beside real matrix-Lie groups complex matrix-Lie groups also can be considered. So, e.g. $SL(n, \mathbb{C})$ is the Lie group of all complex $n \times n$ matrices with determinant $+1$. Similarly there are matrix-Lie groups whose entries are quaternions.

5.3.6.3 Important Applications

1. Rigid Body Movement

1. The group $SE(3)$ is the group of rigid-body motions in the Euclidean space $\mathbb{R}^3$. That is why it is so often applied in control of robots. The 6 independent transformations are defined usually as follows:

1. Translation in x-direction,
2. Translation in y-direction,
3. Translation in z-direction,
4. Rotation about the x-axis,
5. Rotation about the y-axis,
6. Rotation about the z-axis.

These transformations can be represented by 4×4 matrices applied to homogeneous coordinates (see 3.5.4.2, p. 231) in 3 dimensions, i.e. $(x,\, y,\, z)^{\mathrm{T}} \in \mathbb{R}^3$ is represented as a vector $(x,\, y,\, z,\, 1)^{\mathrm{T}}$ with four coordinates (see 3.5.4.2, p. 231).
Matrices corresponding to the transformations 1 until 6 are:

$$\mathbf{M}_1 = \begin{pmatrix} 1&0&0&a \\ 0&1&0&0 \\ 0&0&1&0 \\ 0&0&0&1 \end{pmatrix}, \quad \mathbf{M}_2 = \begin{pmatrix} 1&0&0&0 \\ 0&1&0&b \\ 0&0&1&0 \\ 0&0&0&1 \end{pmatrix}, \quad \mathbf{M}_3 = \begin{pmatrix} 1&0&0&0 \\ 0&1&0&0 \\ 0&0&1&c \\ 0&0&0&1 \end{pmatrix}, \tag{5.145a}$$

$$\mathbf{M}_4 = \begin{pmatrix} 1&0&0&0 \\ 0&\cos\alpha&-\sin\alpha&0 \\ 0&\sin\alpha&\cos\alpha&0 \\ 0&0&0&1 \end{pmatrix}, \mathbf{M}_5 = \begin{pmatrix} \cos\beta&0&\sin\beta&0 \\ 0&1&0&0 \\ -\sin\beta&0&\cos\beta&0 \\ 0&0&0&1 \end{pmatrix}, \mathbf{M}_6 = \begin{pmatrix} \cos\gamma&-\sin\gamma&0&0 \\ \sin\gamma&\cos\gamma&0&0 \\ 0&0&1&0 \\ 0&0&0&1 \end{pmatrix}. \tag{5.145b}$$

The matrices $\mathbf{M}_4$, $\mathbf{M}_5$, $\mathbf{M}_6$ describe the rotations in $\mathbb{R}^3$, consequently $SO(3)$ is a subgroup of $SE(3)$. The group $SE(3)$ acts on $\vec{\mathbf{x}} = (x,\, y,\, z)^{\mathrm{T}} \in \mathbb{R}^3$ with homogeneous coordinates $(\vec{\mathbf{x}},\, 1)^{\mathrm{T}}$ as follows:

$$\begin{pmatrix} \vec{\mathbf{x}}' \\ 1 \end{pmatrix} = \begin{pmatrix} \mathbf{R} & \vec{\mathbf{v}} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \vec{\mathbf{x}} \\ 1 \end{pmatrix} = \begin{pmatrix} \mathbf{R}\vec{\mathbf{x}} + \vec{\mathbf{v}} \\ 1 \end{pmatrix} \tag{5.146}$$

where $\mathbf{R} \in SO(3)$ is a rotation, and $\vec{\mathbf{v}} = (a,\, b,\, c)^{\mathrm{T}}$ is a translation vector.

2. Affine Transformations of 2-Dimensional Space

The group $GA(2)$ of affine transformations of the 2-dimensional space is a 6-dimensional matrix Lie group with the following 6 dimensions:

1. Translation in x-direction,
2. Translation in y-direction,
3. Rotation about the origin,
4. Stretching or shrinking with respect to the origin,
5. Shearing (stretching with resp. to y, with resp. to x),
6. 45°-shearing with respect to 5.

Also these transformations are described by matrices in homogeneous coordinates $(x, y, 1)^T$ for $(x, y)^T \in \mathbb{R}^2$:

$$\mathbf{M}_1 = \begin{pmatrix} 1&0&a \\ 0&1&0 \\ 0&0&1 \end{pmatrix}, \quad \mathbf{M}_2 = \begin{pmatrix} 1&0&0 \\ 0&1&b \\ 0&0&1 \end{pmatrix}, \quad \mathbf{M}_3 = \begin{pmatrix} \cos\alpha&-\sin\alpha&0 \\ \sin\alpha&\cos\alpha&0 \\ 0&0&1 \end{pmatrix}, \tag{5.147a}$$

$$\mathbf{M}_4 = \begin{pmatrix} e^{\tau} & 0 & 0 \\ 0 & e^{\tau} & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \mathbf{M}_5 = \begin{pmatrix} e^{\mu} & 0 & 0 \\ 0 & e^{-\mu} & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \mathbf{M}_6 = \begin{pmatrix} \cosh\nu & \sinh\nu & 0 \\ \sinh\nu & \cosh\nu & 0 \\ 0 & 0 & 1 \end{pmatrix}. \qquad (5.147b)$$

This group has as essential subgroups the translation group, given by $\mathbf{M}_1$ and $\mathbf{M}_2$, the Euclidean group $SE(2)$, given by $\mathbf{M}_1$, $\mathbf{M}_2$ and $\mathbf{M}_3$, the similarity group, given by $\mathbf{M}_1$, $\mathbf{M}_2$, $\mathbf{M}_3$, $\mathbf{M}_4$.

Application: The group $GA(2)$ can be applied to describe all transformations of a planar object which is recorded under slight angle modifications by a camera moving in the 3 dimensional space.
If also large changes in angles of perspectivity can occur, then group $P(2)$ the group of all transformations of projective spaces can be used. The matrix-Lie group is generated by the matrices $\mathbf{M}_1$ until $\mathbf{M}_6$ and by the two further matrices

$$\mathbf{M}_7 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \beta & 0 & 1 \end{pmatrix}, \qquad \mathbf{M}_8 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \gamma & 1 \end{pmatrix}. \qquad (5.147c)$$

These two additional matrices correspond to a change of the horizon or vanishing of an edge of the plane picture.

5.3.6.4 Lie Algebra

1. Real Lie algebra

A *real Lie algebra* $\mathcal{A}$ is a real vector space with an operation

$$[\cdot\,,\,\cdot] \,:\, \mathcal{A} \times \mathcal{A} \to \mathcal{A}, \qquad (5.148)$$

which is called the Lie bracket and for which the following properties are valid for all $a, b, c \in \mathcal{A}$:

- $[.\,,\,.]$ is bilinear,
- $[a\,,\,b] = -[b\,,\,a]$, i.e. the operation is skew-symmetric or anticommutative,
- the so called Jacobi identity is valid (as a replacement of the missing associativity)

$$[a\,,\,[b\,,\,c]] + [c\,,\,[a\,,\,b]] + [b\,,\,[c,\,a]] = 0. \qquad (5.149)$$

Obviously $[a\,,\,a] = 0$ holds.

2. Lie Bracket

For (real) $n \times n$ matrices $\mathbf{X}$ and $\mathbf{Y}$ a Lie bracket is given by the commutator, i.e.

$$[\mathbf{X},\,\mathbf{Y}] := \mathbf{XY} - \mathbf{YX}. \qquad (5.150)$$

3. Special Lie-Algebras

There are associated Lie algebras to matrix-Lie groups.

1. A function $g : \mathbb{R} \to GL(n)$ is a *one-parameter subgroup* of $GL(n)$, if
- g is continuous,
- $g(0) = I_{n\times n}$,
- $g(t+s) = g(t)g(s)$ for every $t, s \in \mathbb{R}$.

In particular:

2. If g is a one-parameter subgroup of $GL(n)$, then there exists a uniquely defined matrix $\mathbf{X}$ such that

$$g(t) = e^{t\mathbf{X}} \quad \text{(see 5.3.6.1, p. 351)}. \qquad (5.151)$$

3. For every $n \times n$ matrix $\mathbf{A}$ the logarithm $\log \mathbf{A}$ is defined by

$$\log A = \sum_{m=1}^{\infty} \frac{(-1)^{m+1}}{m} (A - I)^m\,, \qquad (5.152)$$

if this series is convergent. In particular, the series converges if $||\mathbf{A} - \mathbf{I}|| < 1$.

4. Correspondence between Lie Group and Lie Algebra

The correspondence between a matrix-Lie group and the associated Lie algebra is as follows.

1. Let G be a matrix-Lie group. The Lie *algebra of* G, which is denoted by $\mathbf{g}$, is the set of all matrices $\mathbf{X}$ such that $e^{t\mathbf{X}} \in G$ holds for all real numbers t.

In a given matrix-Lie group the elements close to the unit matrix can be represented as $g(t) = e^{t\mathbf{X}}$ with $\mathbf{X} \in \mathbf{g}$, and t close to zero. If the exponential map is surjective, as in the case of $SO(n)$ and $SE(n)$, then the elements of the group can be parameterized with the help of the matrix-exponential function by elements of the corresponding Lie algebra. The matrices $\frac{dg}{dt}g^{-1}$ and $g^{-1}\frac{dg}{dt}$ respectively are called *tangent vectors* or *tangent elements* to $g \in G$. Calculating these elements for $t = 0$, one gets $\mathbf{X}$ itself, i.e. $\mathbf{g}$ is the tangent space $T_{\mathbf{I}}G$ at the identity matrix $\mathbf{I}$.

2. It can be shown that the Lie algebra assigned to a Lie group in this way is a Lie algebra also in the abstract sense.

Let G be a matrix-Lie group with the associated matrix-Lie algebra $\mathbf{g}$ and $\mathbf{X}$ and $\mathbf{Y}$ elements of $\mathbf{g}$. Then:

- $s\mathbf{X} \in \mathbf{g}$ for any real numbers s,
- $\mathbf{X} + \mathbf{Y} \in \mathbf{g}$,
- $[\mathbf{X},\, \mathbf{Y}] = \mathbf{XY} - \mathbf{YX} \in \mathbf{g}$.

■ **A:** The Lie algebra $\mathbf{so}(2)$ associated to the Lie group $SO(2)$ is calculated from the representation of the elements $g(\theta) = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$ by $SO(2)$ with the help of the tangential elements

$$\frac{dg}{d\theta}g^{-1}\bigg|_{\theta=0} = \begin{pmatrix} -\sin\theta & -\cos\theta \\ \cos\theta & -\sin\theta \end{pmatrix}\begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}\bigg|_{\theta=0} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}. \tag{5.153a}$$

Consequently

$$\mathbf{so}(2) = \left\{ s\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \quad s \in \mathbb{R} \right\}. \tag{5.153b}$$

Conversely, from

$$\mathbf{X} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \quad \text{comes} \quad e^{s\mathbf{X}} = \cos s\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \sin s\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} \cos s & -\sin s \\ \sin s & \cos s \end{pmatrix}. \tag{5.153c}$$

■ **B:** The following matrices form a basis for the Lie algebra $\mathbf{so(3)}$:

$$\mathbf{X}_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}, \quad \mathbf{X}_2 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix}, \quad \mathbf{X}_3 = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{5.154}$$

Remark: The surjectivity of the exponential mappings $\mathbf{so}(3) \to SO(3)$ and $\mathbf{se}(3) \to SE(3)$ implies the existence of a (many-valued) logarithmic function. Nevertheless this logarithm function can be applied to interpolation.

E.g. if rigid-body motions $\mathbf{B}_1,\, \mathbf{B}_2 \in SE(3)$ are given, then $\log \mathbf{B}_1$, $\log \mathbf{B}_2$ can be calculated which are elements of the Lie algebra $\mathbf{so}(3)$. Then between these logarithms linear interpolation $(1-t)\log \mathbf{B}_1 + t\log \mathbf{B}_2$ can be taken and then the exponential map can be applied in order to get an interpolation between the rigid-body motions $\mathbf{B}_1$ and $\mathbf{B}_2$ by

$$\exp\left((1-t)\log \mathbf{B}_1 + t\log \mathbf{B}_2\right). \tag{5.155}$$

■ **C:** The matrix-Lie algebra $\mathbf{se}(3)$ associated to the matrix-Lie group $SE(3)$ is generated by the matrices:

$$\mathbf{E}_1 = \begin{pmatrix} 0&0&0&1 \\ 0&0&0&0 \\ 0&0&0&0 \\ 0&0&0&0 \end{pmatrix}, \quad \mathbf{E}_2 = \begin{pmatrix} 0&0&0&0 \\ 0&0&0&1 \\ 0&0&0&0 \\ 0&0&0&0 \end{pmatrix}, \quad \mathbf{E}_3 = \begin{pmatrix} 0&0&0&0 \\ 0&0&0&0 \\ 0&0&0&1 \\ 0&0&0&0 \end{pmatrix}, \tag{5.156a}$$

$$\mathbf{E}_4 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad \mathbf{E}_5 = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad \mathbf{E}_6 = \begin{pmatrix} 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}. \tag{5.156b}$$

5. Inner Product

For a given finite dimensional matrix-Lie group it is always possible to find an orthonormal basis for the associated Lie algebra if a suitable inner product (scalar product) is defined. In this case from any basis of the Lie algebra an orthonormal basis can be obtained by the Gram-Schmidt orthogonalization process (see 4.6.2.2, **4.** p. 316).

In the case of a real matrix-Lie group the Lie algebra consists of real matrices and so an inner product is given by

$$(\mathbf{X}, \mathbf{Y}) = \frac{1}{2}\mathrm{Spur}\,(\mathbf{X}\mathbf{W}\mathbf{Y}^{\mathrm{T}}) \tag{5.157}$$

with a positive definite real symmetric matrix $\mathbf{W}$.

■ **A:** The group of rigid-body motions $SE(2)$ can be parametrized as

$$g(x_1, x_2, \theta) = e^{x_1\mathbf{X}_1 + x_2\mathbf{X}_2}\, e^{\theta\,\mathbf{X}_3} = \begin{pmatrix} \cos\theta & -\sin\theta & x_1 \\ \sin\theta & \cos\theta & x_2 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{with} \tag{5.158a}$$

$$\mathbf{X}_1 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \mathbf{X}_2 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}, \quad \mathbf{X}_3 = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{5.158b}$$

Here $\mathbf{X}_1$, $\mathbf{X}_2$, $\mathbf{X}_3$ form an orthonormal basis of Lie algebra $\mathbf{se}(2)$ with respect to an inner product given by the weight matrix

$$\mathbf{W} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix} \tag{5.158c}$$

■ **B:** A basis of Lie algebra $\mathbf{sl}(2, \mathbb{R})$ is

$$\mathbf{X}_1 = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \quad \mathbf{X}_2 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \quad \text{and} \quad \mathbf{X}_3 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}. \tag{5.159}$$

These elements form an orthonormal basis with respect to the weight matrix $\mathbf{W} = \mathbf{I}_{2\times 2} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.

5.3.6.5 Applications in Robotics

1. Rigid Body Motion

The special Euclidean group $SE(3)$, which describes the rigid-body motions in $\mathbb{R}^3$, is the semidirect product of group $SO(3)$ (rotation about the origin) and $\mathbb{R}^3$ (translations):

$$SE(3) = SO(3) \times \mathbb{R}^3\,. \tag{5.160}$$

In a direct product the factors have no interaction, but this is a semidirect product since rotations act on translations as it is clear from matrix multiplication:

$$\begin{pmatrix} \mathbf{R}_2 & \vec{\mathbf{t}}_2 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \mathbf{R}_1 & \vec{\mathbf{t}}_1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} \mathbf{R}_2\mathbf{R}_1 & \mathbf{R}_2\vec{\mathbf{t}}_1 + \vec{\mathbf{t}}_2 \\ 0 & 1 \end{pmatrix}, \tag{5.161}$$

i.e. the first translation vector is rotated before the second translation vector is added.

2. Theorem of Chasles

This theorem tells that every rigid-body motion which is not a pure translation can be described as a (finite) screw motion. A (finite) screwing motion along an axis through the origin has the form

$$A(\theta) = \begin{pmatrix} \mathbf{R} & \dfrac{\theta\, p}{2\pi}\vec{\mathbf{x}} \\ 0 & 1 \end{pmatrix}, \tag{5.162a}$$

where $\vec{\mathbf{x}}$ is a unit vector in the direction of the axis of rotation, θ is the angle of rotation and p is the angular coefficient. Since $\vec{\mathbf{x}}$ is the axis of rotation $\mathbf{R}\vec{\mathbf{x}} = \vec{\mathbf{x}}$, i.e. $\vec{\mathbf{x}}$ is an eigenvector of matrix $\mathbf{R}$ belonging to unit eigenvalue 1 .

When the axis of rotation does not go through the origin, then a point $\vec{\mathbf{u}}$ of the axis of rotation is chosen which is shifted into the origin, then after the screwing it is shifted back:

$$\begin{pmatrix} \mathbf{I} & \vec{\mathbf{u}} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \mathbf{R} & \dfrac{\theta\, p}{2\pi}\vec{\mathbf{x}} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \mathbf{I} & -\vec{\mathbf{u}} \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} \mathbf{R} & \dfrac{\theta\, p}{2\pi}\vec{\mathbf{x}} + (\mathbf{I} - \mathbf{R})\vec{\mathbf{u}} \\ 0 & 1 \end{pmatrix}. \tag{5.162b}$$

The theorem of Chasles tells that an arbitrary rigid-body motion can be given in the above form, i.e.

$$\begin{pmatrix} \mathbf{R} & \vec{\mathbf{t}} \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} \mathbf{R} & \dfrac{\theta\, p}{2\pi}\vec{\mathbf{x}} + (\mathbf{I} - \mathbf{R})\vec{\mathbf{u}} \\ 0 & 1 \end{pmatrix} \tag{5.163}$$

for given $\mathbf{R}$, $\vec{\mathbf{t}}$ and appropriate p and $\vec{\mathbf{u}}$. Assuming that the angle of rotation θ and the axis of rotation $\vec{\mathbf{x}}$ are already known from $\mathbf{R}$

$$\frac{\theta\, p}{2\pi} = \vec{\mathbf{x}} \cdot \vec{\mathbf{t}} \tag{5.164}$$

is valid, so the angular coefficient p can be calculated. Then the solution of a linear system of equations gives $\vec{\mathbf{u}}$:

$$(\mathbf{I} - \mathbf{R})\vec{\mathbf{u}} = \frac{\theta\, p}{2\pi}\vec{\mathbf{x}} - \vec{\mathbf{t}}\,. \tag{5.165}$$

This is a singular system of equations, where $\vec{\mathbf{x}}$ is in its kernel. Therefore the solution $\vec{\mathbf{u}}$ is unique except to a manyfold of $\vec{\mathbf{x}}$. In order to determine $\vec{\mathbf{u}}$ it is reasonable to require that $\vec{\mathbf{u}}$ is perpendicular to $\vec{\mathbf{x}}$. When the rigid body motion is a pure rotation, then it is not possible to determine an appropriate vector $\vec{\mathbf{u}}$.

3. Mechanical Joints

Joints with one degree of freedom can be represented by a one-parameter subgroup of the group $SE(3)$. For the general case of screw joints the corresponding subgroup is

$$A(\theta) = \begin{pmatrix} \mathbf{R} & \dfrac{\theta\, p}{2\pi}\vec{\mathbf{x}} + (\mathbf{I} - \mathbf{R})\vec{\mathbf{u}} \\ 0 & 1 \end{pmatrix}, \tag{5.166}$$

where $\vec{\mathbf{x}}$ is the axis of rotation, θ is the angle of rotation, p gives the angular coefficient and $\vec{\mathbf{u}}$ is an arbitrary point on the axis of rotation.

The most often occurring types of joints are the rotational joints which can be described by the following subgroup:

$$A(\theta) = \begin{pmatrix} \mathbf{R} & (\mathbf{I} - \mathbf{R})\vec{\mathbf{u}} \\ 0 & 1 \end{pmatrix}. \tag{5.167}$$

The subgroup corresponding the shift joints is

$$A(\theta) = \begin{pmatrix} \mathbf{I} & \theta\vec{\mathbf{t}} \\ 0 & 1 \end{pmatrix}, \tag{5.168}$$

where $\vec{\mathbf{t}}$ describes the direction of the shifting.

4. Forward Kinematics

The goal in the case of industrial robots is the moving and control of the end effectors, which is done by joints in a kinematic chain. If all joints are of one parameter and the robot consists e.g. of 6 joints, then every position of the robot can be described by the joint-variables $\vec{\theta}^{\mathrm{T}} = (\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6)$. The output state of the robot is described by the null vector. Then the motions of the robot can be described so that first the farest joint together the end effector are moved and this motion is given by the matrix $A(\theta_6)$. Then the 5-th joint is moved. Since the axis of this joint should not be influenced by the motion of the last joint, this motion is given by the matrix $A(\theta_5)$. In this way all the joints are moved, and the complete motion of the end effector is given by

$$K(\vec{\theta}) = A_1(\theta_1)A_2(\theta_2)A_3(\theta_3)A_4(\theta_4)A_5(\theta_5)A_6(\theta_6). \tag{5.169}$$

5. Vector Product and Lie Algebra

A screw is given by

$$A(\theta) = \begin{pmatrix} \mathbf{R} & \frac{\theta p}{2\pi}\vec{\mathbf{x}} + (\mathbf{I} - \mathbf{R})\vec{\mathbf{u}} \\ 0 & 1 \end{pmatrix}; \tag{5.170}$$

and it represents rigid body motions parameterized by the angle θ. Obviously, $\theta = 0$ gives the identity transformation. If the derivative is calculated at $\theta = 0$, i.e. the derivative at the identity, then the general element of the Lie algebra is the following:

$$S = \left.\frac{dA}{d\theta}\right|_{\theta=0} = \left.\begin{pmatrix} \frac{d\mathbf{R}}{d\theta} & \frac{p}{2\pi}\vec{\mathbf{x}} - \frac{d\mathbf{R}}{d\theta}\vec{\mathbf{u}} \\ 0 & 0 \end{pmatrix}\right|_{\theta=0} = \begin{pmatrix} \mathbf{\Omega} & \frac{p}{2\pi}\vec{\mathbf{x}} - \mathbf{\Omega}\vec{\mathbf{u}} \\ 0 & 0 \end{pmatrix}, \tag{5.171a}$$

where $\mathbf{\Omega} = \frac{d\mathbf{R}}{d\theta}(0)$ is a skew symmetric matrix. It can be shown that $\mathbf{R}$ is an orthogonal matrix, so $\mathbf{R}\mathbf{R}^{\mathrm{T}} = \mathbf{I}$ and $\mathbf{R}\mathbf{R}^{\mathrm{T}} = \mathbf{I}$ holds and therefore

$$\frac{d}{d\theta}(\mathbf{R}\,\mathbf{R}^{\mathrm{T}}) = \frac{d\mathbf{R}}{d\theta}\mathbf{R}^{\mathrm{T}} + \mathbf{R}\frac{d\mathbf{R}^{\mathrm{T}}}{d\theta} = \frac{d\mathbf{I}}{d\theta} = 0\,. \tag{5.171b}$$

Since $\mathbf{R} = \mathbf{I}$ for $\theta = 0$

$$\frac{d\mathbf{R}}{d\theta}(0) + \frac{d\mathbf{R}^{\mathrm{T}}}{d\theta}(0) = 0\,. \tag{5.171c}$$

So every skew symmetric matrix

$$\mathbf{\Omega} = \begin{pmatrix} 0 & -\omega_z & \omega_y \\ \omega_z & 0 & -\omega_x \\ -\omega_y & \omega_x & 0 \end{pmatrix} \tag{5.171d}$$

can be identified with a vector $\vec{\boldsymbol{\omega}}^{\mathrm{T}} = (\omega_x, \omega_y, \omega_z)$. In this way the multiplication of any three dimensional vector $\vec{\mathbf{p}}$ by matrix $\mathbf{\Omega}$ corresponds to the vector product with vector $\vec{\boldsymbol{\omega}}$:

$$\mathbf{\Omega}\,\vec{\mathbf{p}} = \vec{\boldsymbol{\omega}} \times \vec{\mathbf{p}}\,. \tag{5.171e}$$

Consequently $\vec{\boldsymbol{\omega}}$ is the angular velocity of the rigid body with an amplitude ω.
Hence a general element of the Lie algebra $\mathbf{se}(3)$ has the form

$$= \begin{pmatrix} \mathbf{\Omega} & \vec{\mathbf{v}} \\ 0 & 0 \end{pmatrix}. \tag{5.171f}$$

These matrices form a 6-dimensional vector space which is often identified with the 6-dimensional vectors of the form

$$\vec{\mathbf{s}} = \begin{pmatrix} \vec{\boldsymbol{\omega}} \\ \vec{\mathbf{v}} \end{pmatrix}. \tag{5.172}$$

5.3.7 Rings and Fields

In this section, there are discussed algebraic structures with two binary operations.

5.3.7.1 Definitions

1. Rings

A set R with two binary operations $+$ and $*$ is called a *ring* (notation: $(R, +, *)$), if

- $(R, +)$ is an Abelian group,
- $(R, *)$ is a semigroup, and
- the *distributive laws* hold:

$$a * (b + c) = (a * b) + (a * c), \quad (b + c) * a = (b * a) + (c * a). \tag{5.173}$$

If $(R, *)$ is commutative or if $(R, *)$ has a neutral element, then $(R, +, *)$ is called a commutative ring or a ring with identity (ring with unit element), respectively.
A commutative ring with a unit element and without zero divisor is called the *domain of integrity*.
A nonzero element of a ring is called *zero divisor* or *singular element* if there is a nonzero element of the ring such that their product is equal to zero.
In a ring with zero divisor the following implication is generally false: $a * b = 0 \implies (a = 0 \vee b = 0)$.
If R is a ring with a unit element, then the *characteristic of the ring* R is the smallest natural number k such that $k1 = 1 + 1 + \ldots + 1 = 0$ (k times 1 equals to zero), and it is denoted by char $R = k$. If such a k does not exist, then char $R = 0$.
char $R = k$ means that the cyclic subgroup $\langle 1 \rangle$ of the additive group $(R, +)$ generated by 1 has order k, so the order of every element is a divisor of k.
If char $R = k$ and for all $r \in R$, then $r + r + \ldots + r$ (k times)is equal to zero. The characteristic of a domain of integrity is zero or a prime.

2. Division Ring, Field

A ring is called *division ring* or *skew field* if $(R \setminus \{0\}, *)$ is a group .
If $(R\setminus\{0\}, *)$ is commutative, then R is a *field*. So, every field is a domain of integrity and also a division ring. Reversed, every finite domain of integrity and every finite division ring is a field. This statement is a theorem of Wedderburn.

Examples of rings and fields

■ **A:** The number domains $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$ are commutative rings with identity with respect to addition and multiplication; $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$ are also fields. The set of even integers is an example of a ring without identity.

■ **B:** The set M_n of all square matrices of order n with real (or complex) elements is a non-commutative ring with respect to matrix addition and multiplication. It has a unit element which is the identity matrix. M_n has zero divisors, e.g. for $n = 2$, $\begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$, i.e. both matrices $\begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}$ and $\begin{pmatrix} 0 & 0 \\ 1 & 1 \end{pmatrix}$ are zero divisors in M_2.

■ **C:** The set of real polynomials $p(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0$ forms a ring with respect to the usual addition and multiplication of polynomials, the polynomial ring $\mathbb{R}[x]$.
More generally, instead of polynomials over $\mathbb{R}$, polynomial rings over arbitrary commutative rings with identity element can be considered.

■ **D:** Examples of finite rings are the *residue class rings* $\mathbb{Z}_n$ modulo n. $\mathbb{Z}_n$ consists of all the classes $[a]_n$ of integers having the same residue on division by n. ($[a]_n$ is the equivalence class defined by the natural number a with respect to the relation $\sim_R$ introduced in 5.2.4, **1.**, p. 334.) The ring operations $\oplus$, $\odot$ on $\mathbb{Z}_n$ are defined by

$$[a]_n \oplus [b]_n = [a + b]_n \quad \text{and} \quad [a]_n \odot [b]_n = [a \cdot b]_n. \tag{5.174}$$

If the natural number n is a prime, then $(\mathbb{Z}_n , \oplus , \odot)$ is a field. Otherwise $\mathbb{Z}_n$ has zero divisors, e.g. in $\mathbb{Z}_6$ (numbers modulo 6) $[3]_6 \cdot [2]_6 = [0]_6$. Usually $\mathbb{Z}_n$ is considered as $\mathbb{Z}_n = \{0, 1, \ldots, n - 1\}$, i.e. the

residue classes are replaced by representatives(see 5.4.3,**3.**, S. 377).

3. Field Extensions

If K and L are fields and $K \subseteq L$, then L is an *extension field* or an *over-field* of K. In this case L can be considered as a vector space over K.

If L is a finite dimensional space over K, then L is called a *finite extension field.* If this dimension is n, then L is called also an *extension of degree* n of K (Notation: $[L : K] = n$).

E.g. $\mathbb{C}$ is a finite extension of $\mathbb{R}$. $\mathbb{C}$ is two-dimensional over $\mathbb{R}$, and $\{1, i\}$ is a basis. $\mathbb{R}$ is an infinite-dimensional space over $\mathbb{Q}$.

For a set $M \subseteq L$, $K(M)$ denotes the smallest field (an over-field of K) which contains the field K and the set M.

Especially important are the simple algebraic extensions $K(\alpha)$, where $\alpha \in L$ is a root of a polynomial from $K[x]$. The polynomial of lowest degree with a leading coefficient 1 having α as a root is called the *minimal polynomial of* α *over* K. If the degree of the minimal polynomials of $\alpha \in L$ is n, then $K(\alpha)$ is an extension of degree n, i.e. the degree of the minimal polynomials is equal to the dimension of L as a vector space over K.

E.g. $\mathbb{C} = \mathbb{R}(i)$ and $i \in \mathbb{C}$ is the root of the polynomial $x^2 + 1 \in \mathbb{R}[x]$, i.e. $\mathbb{C}$ is a simple algebraic extension and $[\mathbb{C} : \mathbb{R}] = 2$.

A field, which does not have any proper subfield, is called a *prime field.*
Every field K contains a smallest subfield, the prime field of K.
Out of isomorphism, $\mathbb{Q}$ (for fields of characteristic 0) and $\mathbb{Z}_p$ (p prime, for fields of characteristic p) are the single prime fields.

5.3.7.2 Subrings, Ideals

1. Subring

Suppose $R = (R, +, *)$ is a ring and $U \subseteq R$. If U with respect to $+$ and $*$ is also a ring, then $U = (U, +, *)$ is called a *subring* of R.
A non-empty subset U of a ring $(R, +, *)$ forms a subring of R if and only if for all $a, b \in U$ also $a + (-b)$ and $a * b$ are in U (subring criterion).

2. Ideal

A subring I is called an *ideal* if for all $r \in R$ and $a \in I$ also $r * a$ and $a * r$ are in I. These special subrings are the basis for the formation of factor rings (see 5.3.7.3, p. 363).
The *trivial subrings* $\{0\}$ and R are always ideals of R. Fields have only trivial ideals.

3. Principal Ideal

If all the elements of an ideal can be generated by one element according to the subring criterion, then it is called a *principal ideal.* All ideals of $\mathbb{Z}$ are principal ideals. They can be written in the form $m\mathbb{Z} = \{mg | g \in \mathbb{Z}\}$ and they are denoted by (m).

5.3.7.3 Homomorphism, Isomorphism, Homomorphism Theorem

1. Ring Homomorphism and Ring Isomorphism

1. Ring Homomorphism: Let $R_1 = (R_1, +, *)$ and $R_2 = (R_2, \circ_+, \circ_*)$ be two rings. A mapping $h\colon R_1 \to R_2$ is called a *ring homomorphism* if for all $a, b \in R_1$

$$h(a + b) = h(a) \circ_+ h(b) \quad \text{and} \quad h(a * b) = h(a) \circ_* h(b) \tag{5.175}$$

hold.

2. Kernel: The *kernel* of h is the set of elements of R_1 whose image by h is the neutral element 0 of $(R_2, +)$, and it is denoted by $\ker h$:

$$\ker h = \{a \in R_1 | h(a) = 0\}. \tag{5.176}$$

Here $\ker h$ is an ideal of R_1.

3. Ring Isomorphism: If h is also bijective, then h is called a *ring isomorphism*, and the rings R_1 and R_2 are called isomorphic.

4. Factor Ring: If I is an ideal of a ring $(R, +, *)$, then the sets of co-sets $\{a + I|a \in R\}$ of I in the additive group $(R, +)$ of the ring R (see 5.3.3, **1.**, p. 337) form a ring with respect to the operations

$$(a + I) \circ_+ (b + I) = (a + b) + I \quad \text{and} \quad (a + I) \circ_* (b + I) = (a * b) + I. \qquad (5.177)$$

This ring is called the *factor ring* of R by I, and it is denoted by R/I.
The factor ring of Z by a principal ideal (m) is the residue class ring $Z_m = Z_{/(m)}$ (see examples of rings and fields on p. 361).

2. Homomorphism Theorem for Rings

If the notion of a normal subgroup is replaced by the notion of an ideal in the homomorphism theorem for groups, then the *homomorphism theorem for rings* is obtained: A ring homomorphism h: $R_1 \to R_2$ defines an ideal of R_1, namely $\ker h = \{a \in R_1|h(a) = 0\}$. The factor ring $R_1/\ker h$ is isomorphic to the homomorphic image $h(R_1) = \{h(a)|a \in R_1\}$. Conversely, every ideal I of R_1 defines a homomorphic mapping nat_I: $R_1 \to R_2/I$ with $nat_I(a) = a+I$. This mapping nat_I is called a *natural homomorphism*.

5.3.7.4 Finite Fields and Shift Registers

1. Finite Fields

The following statements give an overview of the structure of finite fields.

1. Galois Field GF For every power of primes p^n there exits a unique field with p^n elements (out of an isomorphism), and every finite field has p^n elements. The fields with p^n elements are denoted by GF(p^n) (Galois field).
Note: For $n > 1$ GF(p^n) and $\mathbb{Z}_{p^n}$ are different.

In constructing finite fields with p^n elements (p is prime, $n > 1$), the ring of polynomials over $\mathbb{Z}_p$ (see 5.3.7,**2.**, p. 361, ■ **C**) and irreducible polynomials are needed: $\mathbb{Z}_p[x]$ consists of all polynomials with coefficients from $\mathbb{Z}_p$. The coefficients are calculated modulo p.

2. Algorithm of Division and Euclidean Algorithm In a ring of polynomials $K[x]$ the division algorithm is applicable (dividing polynomials with a remainder), i.e. for $f(x), g(x) \in K[x]$, degf(x) $\le$ degg(x) there exist $q(x), r(x) \in K[x]$ such that

$$g(x) = q(x) \cdot f(x) + r(x) \text{ and } \deg r(x) < \deg f(x)\,. \qquad (5.178)$$

This relation is denoted by $r(x) = g(x)(\text{mod} f(x))$. Repeatedly performed division with remainders is known as the Euclidean algorithm for rings of polynomials and the last nonzero remainder gives the greatest common divisor of $f(x)$ and $g(x)$.

3. Irreducible Polynomials A polynomial $f(x) \in K[x]$ is *irreducible* if it can not be represented as a product of polynomials of lower degrees, i.e. (analogously to the prime numbers in Z) $f(x)$ is a prime in $K[x]$. E.g. for polynomials of second or third degree irreducibility means, that they do not have roots in K.

It can be shown that there are irreducible polynomials of arbitrary degree in $K[x]$. If $f(x) \in K[x]$ is an irreducible polynomial, then

$$K[x]/f(x) := \{p(x) \in K[x] \mid \deg p(x) < \deg f(x)\} \qquad (5.179)$$

is a field, where addition and multiplication are performed modulo $f(x)$, i.e. $g(x) * h(x) = g(x) \cdot h(x)\,(\text{mod}\, f(x))$.

If $K = \mathbb{Z}_p$ and $\deg f(x) = n$, then $K[x]/f(x)$ has p^n elements, i.e. GF$(p^n) = \mathbb{Z}_p[x]/f(x)$, where $f(x)$ is an irreducible polynomial of degree n.

4. Calculation Rule in GF(p^n) In GF(p^n) the following useful rule is valid:

$$(a + b)^{p^r} = a^{p^r} + b^{p^r}, r \in \mathbb{N}\,. \qquad (5.180)$$

So, in GF(p^n) $=\mathbb{Z}_p[x]/f(x)$ there is an element $\alpha = x$, a root of the polynomial $f(x)$ irreducible in $\mathbb{Z}_p(x)$, and GF$(p^n) = \mathbb{Z}_p[x]/f(x) = \mathbb{Z}_p(\alpha)$. It can be proven that $\mathbb{Z}_p(\alpha)$ is the splitting field of $f(x)$.
The *splitting field* of a polynomial from $\mathbb{Z}_p[x]$ is the smallest extension field of $\mathbb{Z}_p$ which contains all roots of $f(x)$.

5. Algebraic Closure, Fundamental Theorem of Algebra A field K is *algebraically closed* if all roots of the polynomials from $K[x]$ are in K. The *fundamental theorem of algebra* tells that the field $\mathbb{C}$ of complex numbers is algebraically closed. An algebraic extension L of K is called the *algebraic closure* of K if L is algebraically closed. The algebraic closure of a finite field is not finite. So there are infinite fields with characteristic p.

6. Cyclic and Multiplicative Group The multiplicative group $K^* = K \setminus \{0\}$ of a finite field K is cyclic, i.e. there is an element $a \in K$ such that every element of K^* is a power of a: $K^* = \{1, a, a^2, \ldots, a^{q-2}\}$, if K has q elements.

An irreducible polynomial $f(x) \in K[x]$ is called *primitive*, if the powers of x represents all nonzero elements of $L := K[x]/f(x)$, i.e. the multiplicative group L^* of L can be generated by x.

With a primitive polynomial $f(x)$ of degree n it is possible to construct a ,,Table of logarithm" for $\mathrm{GF}(p^n)$ from $\mathrm{GF}(p)[x]$, which makes calculations easier.

■ Construction of field $\mathrm{GF}(2^3)$ and its table of logarithm.

$f(x) = 1 + x + x^3$ is irreducible over $\mathbb{Z}[x]$, since neither 0 nor 1 are roots of it:

$$\mathrm{GF}(2^3) = \mathbb{Z}_2[x]/f(x) = \{a_0 + a_1x + a_2x^2 \mid a_0, a_1, a_2 \in \mathbb{Z}_2 \wedge x^3 = 1 + x\}. \tag{5.181}$$

$f(x)$ is primitive, so a table of logarithm can be created for $\mathrm{GF}(2^3)$:

Two expressions are assigned to every polynomial $a_0 + a_1x + a_2x^2$ from $\mathbb{Z}_2[x]/f(x)$. The coefficient vector a_0, a_1, a_2 and the so called logarithm which is a natural number i such that $x^i = a_0 + a_1x + a_2x^2$ modulo $1 + x + x^3$. The table of logarithm is:

KE	KV	Log.
1	1 0 0	0
x	0 1 0	1
x^2	0 0 1	2
x^3	1 1 0	3
x^4	0 1 1	4
x^5	1 1 1	5
x^6	1 0 1	6

- Addition of the field elements (KE) in GF(8):
- Addition of the coordinate vectors (KV) componentwise mod 2 (in general mod p).
- Multiplication of the field elements (KE) in GF(8):
- Addition of logarithms (Log) mod 7 (in general mod $(p^n - 1)$).

Example: $\dfrac{x^2 + x^4}{x^3 + x^4} = \dfrac{x}{x^6} = x^{-5} = x^2$

Remark: Finite fields are extremely important in *coding theory* as *linear codes*, where vector spaces in form $(\mathrm{GF}(q))^n$ are considered. A subspace of such a vector space is called *linear code* (see 5.4.6.2,**3.**, p. 385). The elements (code words) of a linear code are also n-tuples with elements from a finite field $\mathrm{GF}(q^n)$. In applications in code theory it is important to know the divisors of $X^n - 1$.

The splitting field of $X^n - 1 \in K[X]$ is called the *n-th cyclotomic field* over K.

If the characteristic of K is not a divisor of n and α is a primitive n-th unit root, then:

a) The extension field $K(\alpha)$ is the splitting field of $X^n - 1$ over K.

b) In $K(\alpha)$, the field $X^n - 1$ has exactly n pairwise different roots which form a cyclic group, and among them there are $\varphi(n)$ primitive n-th unit roots, where $\varphi(n)$ denotes the Euler function (5.4.4,**1.**, p. 381). By the k-th powers ($k < n$, g.c.d.(k,n)=1) of a primitive n-th unit root α all unit roots can be got.

2. Applications of Shift Registers

Calculations with polynomials can be performed well by a *linear feedback shift register* (see **Fig.5.18**). With a linear feedback shift register based on the feedback polynomial $f(x) = f_0 + f_1x + \cdots + f_{r-1}x^{r-1} + x^r$ and from the state polynomial $s(x) = s_0 + s_1x + \cdots + s_{r-1}x^{r-1}$ one gets the state polynomial $s(x) \cdot x - s_{r-1} \cdot f(x) = s(x) \cdot x \pmod{f(x)}$.

Especially, if $s(x) = 1$, after i steps (i-times applications) the state polynomial is $x^i \pmod{f(x)}$.

■ Demonstration with the example from page 364: The primitive polynomial $f(x) = 1 + x + x^3 \in \mathbb{Z}_2[x]$ is chosen as feedback polynomial. Then the shift register with lengh 3 has the following sequence of states:

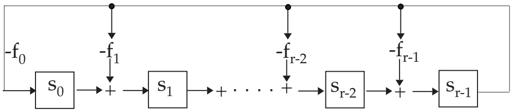


Figure 5.18

From the initial state:	$1\,0\,0 \mathrel{\hat{=}} 1$		$(\text{mod } f(x))$
the states follow as:	$0\,1\,0 \mathrel{\hat{=}} x$		$(\text{mod } f(x))$
	$0\,0\,1 \mathrel{\hat{=}} x^2$		$(\text{mod} f(x))$
	$1\,1\,0 \mathrel{\hat{=}} x^3$	$\equiv 1 + x$	$(\text{mod } f(x))$
	$0\,1\,1 \mathrel{\hat{=}} x^4$	$\equiv x + x^2$	$(\text{mod } f(x))$
	$1\,1\,1 \mathrel{\hat{=}} x^5$	$\equiv 1 + x + x^2$	$(\text{mod } f(x))$
	$1\,0\,1 \mathrel{\hat{=}} x^6$	$\equiv 1 + x^2$	$(\text{mod } f(x))$
	$1\,0\,0 \mathrel{\hat{=}} x^7$	$\equiv 1$	$(\text{mod } f(x))$

The states are considered as coefficient vectors of a state polynomial $s_0 + s_1 x + s_2 x^2$.
In general: A linear feedback shift register with length r gives a sequence of states of maximal length with period $2^r - 1$ if and only if the feedback polynomial is a primitive polynomial of degree r.

5.3.8 Vector Spaces *

5.3.8.1 Definition

A *vector space* over a field F consists of an Abelian group $V = (V, +)$ of " vectors " written in additive form, of a field $F = (F, +, *)$ of " scalars " and an exterior multiplication $F \times V \to V$, which assigns to every ordered pair (k, v) for $k \in F$ and $v \in V$ a vector $kv \in V$. These operations have the following properties:

(V1) $(u + v) + w = u + (v + w)$ for all $u, v, w \in V$. (5.182)

(V2) There is a vector $0 \in V$ such that $v + 0 = v$ for every $v \in V$. (5.183)

(V3) To every vector v there is a vector $-v$ such that $v + (-v) = 0$. (5.184)

(V4) $v + w = w + v$ for every $v, w \in V$. (5.185)

(V5) $1v = v$ for every $v \in V$, 1 denotes the unit element of F. (5.186)

(V6) $r(sv) = (rs)v$ for every $r, s \in F$ and every $v \in V$. (5.187)

(V7) $(r + s)v = rv + sv$ for every $r, s \in F$ and every $v \in V$. (5.188)

(V8) $r(v + w) = rv + rw$ for every $r \in F$ and every $v, w \in V$. (5.189)

If $F = \mathbb{R}$ holds, then it is called a *real vector space.*

Examples of vector spaces:

■ **A:** Single-column or single-row real matrices of type $(n, 1)$ and $(1, n)$, respectively, with respect to matrix addition and exterior multiplication with real numbers form real vector spaces $\mathbb{R}^n$ (the vector space of column or row vectors; see also 4.1.3, p. 271).

■ **B:** All real matrices of type (m, n) form a real vector space.

■ **C:** All real functions continuous on an interval $[a, b]$ with the operations

$$(f + g)(x) = f(x) + g(x) \quad \text{and} \quad (kf)(x) = k \cdot f(x) \tag{5.190}$$

*In this paragraph, generally, vectors are not printed in bold face.

form a real vector space.
Function spaces have a fundamental role in functional analysis (see Ch. 12, p. 654). For further examples see 12.1.2, p. 655.

5.3.8.2 Linear Dependence

Let V be a vector space over F. The vectors $v_1, v_2, \ldots, v_m \in$ V are called *linearly dependent* if there are $k_1, k_2, \ldots, k_m \in K$ not all of them equal to zero such that $0 = k_1v_1 + k_2v_2 + \cdots + k_mv_m$ holds. Otherwise they are *linearly independent*. Linear dependence of at least two vectors means that one of them is a multiple of the other.

If there is a maximal number n of linearly independent vectors in a vector space V, then the vector space V is called *n dimensional*. This number n is uniquely defined and it is called the *dimension*. Every n linearly independent vectors of V form a *basis*. If such a maximal number does not exist, then the vector space is called *infinite dimensional*. The vector spaces in the above examples are n, $m \cdot n$, and infinite dimensional.
In the vector space $\mathbb{R}^n$, n vectors are independent if and only if the determinant of the matrix, whose columns or rows are these vectors, is not equal to zero.
If $\{v_1, v_2, \ldots, v_n\}$ form a basis of an n-dimensional vector space over F, then every vector $v \in V$ has a *unique* representation $v = k_1v_1 + k_2v_2 + \cdots + k_nv_n$ with $k_1, k_2 \ldots, k_n \in F$.
Every set of linearly independent vectors can be completed into a basis of the vector space.

5.3.8.3 Linear Operators

1. Definition of Linear Operators

Let V and W be two real vector spaces. A mapping $f : V \longrightarrow W$ from V into W is called a *linear mapping* or *linear transformation* or *linear operator* (see also 12.1.5.2, p. 658) from V into W if

$$f(u+v) = fu + fv \quad \text{for all } u, v \in V, \tag{5.191}$$

$$f(\lambda u) = \lambda fu \quad \text{for all } u \in V \text{ and all real } \lambda. \tag{5.192}$$

■ **A:** The mapping $fu := \int\limits_{\alpha}^{\beta} u(t)\,dt$, which transforms the space $\mathcal{C}[\alpha, \beta]$ of continuous real functions into the space of real numbers is linear.
In the special case when $W = \mathbb{R}^1$, as in the previous example, linear transformations are called *linear functionals*.
■ **B:** Let $V = \mathbb{R}^n$ and let W be the space of all real polynomials of degree at most $n-1$. Then the mapping $f(a_0, a_1, \ldots, a_{n-1}) := a_0 + a_1x + a_2x^2 + \cdots + a_{n-1}x^{n-1}$ is linear. In this case each n-element vector corresponds to a polynomial of degree $\le n-1$.
■ **C:** If $V = \mathbb{R}^n$ and $W = \mathbb{R}^m$, then all linear operators f from V into W ($f : \mathbb{R}^n \longrightarrow \mathbb{R}^m$) can be characterized by a real matrix $\mathbf{A} = (a_{ik})$ of type (m, n). The relation $\mathbf{A}\underline{\mathbf{x}} = \underline{\mathbf{y}}$ corresponds to the system of linear equations (4.174a)

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & & & \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}.$$

2. Sum and Product of Two Linear Operators

Let $f\colon V \longrightarrow W$, $g\colon V \longrightarrow W$ and $h\colon W \longrightarrow U$ be linear operators. Then the

sum $f + g\colon V \longrightarrow W$ is defined as $(f+g)u = fu + gu$ for all $u \in V$ and the (5.193)

product $hf\colon V \longrightarrow U$ is defined as $(hf)u = h(fu)$ for all $u \in V$. (5.194)

Remarks:
1. If f, g and h are linear, then $f + g$ and fh are also linear operators.

2. The product (5.194) of two linear operators represents the consecutive application of these operators f and h.

3. The product of two linear operators is usually non-commutative even if the products exist:

$$hf \neq fh\,. \tag{5.195a}$$

Commutability exists, if

$$hf - fh = 0 \tag{5.195b}$$

holds. In quantum mechanics the left-hand side of this equation $hf - fh$ is called the *commutator*. In the case (5.195a) the operators f and h do not commutate, therefore we have to be very careful about the order.

■ As a particular example of sums and products of linear operators one may think of sums and products of the corresponding real matrices.

5.3.8.4 Subspaces, Dimension Formula

1. Subspace: Let V be a vector space and U a subset of V. If U is also a vector space with respect to the operations of V, then U is called a *subspace* of V.
A non-empty subset U of V is a subspace if and only if for every $u_1, u_2 \in U$ and every $k \in F$ also $u_1 + u_2$ and $k \cdot u_1$ are in U (*subspace criterion*).

2. Kernel, Image: Let V_1, V_2 be vector spaces over F. If f: $V_1 \to V_2$ is a linear mapping, then the linear subspaces *kernel* (notation: ker f) and *image* (notation: im f) are defined in the following way:

$$\ker f = \{v \in V | f(v) = 0\}, \quad \text{im } f = \{f(v) | v \in V\}. \tag{5.196}$$

So, for example, the solution set of a homogeneous linear equation system $\mathbf{A}\underline{\mathbf{x}} = \underline{\mathbf{0}}$ is the kernel of the linear mapping defined by the coefficient matrix $\mathbf{A}$.

3. Dimension: The dimension dim ker f and dim im f are called the *defect* f and *rank* f, respectively. For these dimensions the equality

$$\text{defect } f + \text{rank } f = \dim V, \tag{5.197}$$

is valid and is called the *dimension formula*. In particular, if the defect $f = 0$, i.e., $\ker f = \{0\}$, then the linear mapping f is injective, and conversely. Injective linear mappings are called *regular*.

5.3.8.5 Euclidean Vector Spaces, Euclidean Norm

In order to be able to use notions such as length, angle, orthogonality in abstract vector spaces we introduce *Euclidean vector spaces*.

1. Euclidean Vector Space

Let V be a real vector space. If φ: $V \times V \to \mathbb{R}$ is a mapping with the following properties (instead of $\varphi(v, w)$ one writes $v \cdot w$) for every $u, v, w \in V$ and for every $r \in \mathbb{R}$

(S1) $v \cdot w = w \cdot v,$ (5.198)

(S2) $(u + v) \cdot w = u \cdot w + v \cdot w,$ (5.199)

(S3) $r(v \cdot w) = (rv) \cdot w = v \cdot (rw),$ (5.200)

(S4) $v \cdot v > 0$ if and only if $v \neq 0,$ (5.201)

then φ is called a *scalar product* on V. If there is a scalar product defined on V, then V is called a *Euclidean vector space*.

These properties are used to define a scalar product with similar properties on more general spaces, too (see 12.4.1.1, p. 673).

2. Euclidean Norm

The value

$$\|v\| = \sqrt{v \cdot v} \tag{5.202}$$

denotes the *Euclidean norm* (length) of v. The angle α between v, w from V is defined by the formula

$$\cos\alpha = \frac{v \cdot w}{\|v\| \cdot \|w\|}\,. \tag{5.203}$$

If $v \cdot w = 0$ holds, then v and w are called *orthogonal* to each other.

■ **Orthogonality of Trigonometric Functions:** In the theory of Fourier series (see 7.4.1.1, p. 474), there are functions of the form $\sin kx$ and $\cos kx$. Theses fubctions can be considered as elements of $C[0, 2\pi]$. In the function space $C[a, b]$ the formula

$$f \cdot g = \int_a^b f(x)g(x)\,dx \tag{5.204}$$

defines a scalar product. Since

$$\int_0^{2\pi} \sin kx \cdot \sin lx\,dx = 0 \quad (k \neq l), \tag{5.205}$$ $$\int_0^{2\pi} \cos kx \cdot \cos lx\,dx = 0 \quad (k \neq l), \tag{5.206}$$

$$\int_0^{2\pi} \sin kx \cdot \cos lx\,dx = 0, \tag{5.207}$$

the functions $\sin kx \in \mathbb{C}[0, 2\pi]$ and $\cos lx \in \mathbb{C}[0, 2\pi]$ for every $k, l \in \mathbb{N}$ are pairwise orthogonal to each other. This *orthogonality of trigonometric functions* is used in the calculation of Fourier coefficients in harmonic analysis (see 7.4.1.1, p. 474).

5.3.8.6 Bilinear Mappings, Bilinear Forms

Bilinear mappings can be considered as generalizations of different products between vectors. In that case bilinearity uses the distributivity of the corresponding product with respect to vector addition.

1. Definition

Let U, V, W be vector spaces over the same field K. A mapping $f\colon U \times V \longrightarrow W$ is called *bilinear* if

for every $u \in U$ the mapping $v \mapsto f(u, v)$ is a linear mapping of V into W and
for every $v \in V$ the mapping $u \mapsto f(u, v)$ is a linear mapping of U into W. (5.208)

It means that a mapping $f\colon U \times V \longrightarrow W$ is bilinear, if for every $k \in K$, $u, u' \in U$, and $v, v' \in V$ holds:

$$f(u + u', v) = f(u, v) + f(u', v)\,,\quad f(ku, v) = kf(u, v) \text{ and}$$
$$f(u, v + v') = f(u, v) + f(u, v')\,,\quad f(u, kv) = kf(u, v)\,. \tag{5.209}$$

If f is replaced by the dot product or vector product or by a multiplication in a field, these relations describe the left sided and right sided distributivity of this multiplication with respect to vector addition.

Especially, if $U = V$, and $W = K$ which is the underlying field, then f is called a *bilinear form.* In this book only the real ($K = \mathbb{R}$) or complex ($K = \mathbb{C}$) cases are considered.

Examles of Bilinearforms

■ **A:** $U = V = \mathbb{R}^n$, $W = \mathbb{R}$, f is the dot product in $\mathbb{R}^n$: $f(u, v) = u^{\mathrm{T}}v = \sum_{i=1}^n u_i v_i$, where u_i and v_i $(i = 1, 2, \ldots, n)$ denote the Cartesian coordinates of u and v.

■ **B:** $U = V = W = \mathbb{R}^3$, f is the cross product in $\mathbb{R}^3$:
$f(u, v) = u \times v = (u_2v_3 - v_2u_3\,,\ v_1u_3 - u_1v_3\,,\ u_1v_2 - v_1u_2)^{\mathrm{T}}$.

2. Special Bilinear Forms

A bilinear form $f\colon V \times V \longrightarrow \mathbb{R}$ is called

- symmetric, if $f(v, v') = f(v', v)$ for every $v, v' \in V$,
- skew-symmetric, if $f(v, v') = -f(v', v)$ for every $v, v' \in V$ and
- positive definite, if $f(v, v) > 0$ for every $v \in V$ $v \neq 0$.

So an Euclidean dot product in V (see 5.3.8.5, p. 367) can be characterized as a symmetric, positive definite bilinear form. The canonical Euclidean dot product in $\mathbb{R}^n$ is defined as $f(u, v) = u^{\mathrm{T}}v$.

In finite dimensional spaces V a bilinear form can be represented by a matrix: If $f := V \times V \longrightarrow \mathbb{R}$ is a bilinear form, and $B = (b_1, b_2, \ldots, b_n)$ is a basis of V, then the matrix

$$\mathbf{A}_B(f) = (f(b_i, b_j)_{i,j}) \qquad (5.210)$$

is the *representation matrix* of f with respect to basis B. The bilinear form then can be written in matrix product form:

$$f(v, v') = v^{\mathrm{T}}\mathbf{A}_B(f)v', \qquad (5.211)$$

where v and v' are given with respect to basis B.
The representation matrix is symmetric, if the bilinear form is symmetric. In complex vector spaces (because z^2 can be a negative number) symmetric, positive definite bilinear forms do not have much sense. To define an unitary dot product and also distances and angles with it instead of bilinear form the concept of the so called sesquilinear form is used [5.6], [5.12].

3. Sesquilinear Form

A mapping $f: V \times V \longrightarrow \mathbb{C}$ is called *sesquilinear form* if for every $v, v' \in V$ and $k \in \mathbb{C}$:

$$f(u + u', v) = f(u, v) + f(u', v), \quad f(ku, v) = kf(u, v) \text{ and}$$
$$f(u, v + v') = f(u, v) + f(u, v'), \quad f(u, kv) = k^* f(u, v). \qquad (5.212)$$

where k^* denotes the complex conjugate of k. The function is linear in the first argument and ,,semi-linear" in the second argument. Analogously to the real case ,,symmetry" is defined in the following way:
A sesquilinear form $f: V \times V \longrightarrow \mathbb{C}$ is called *hermitian* if $f(v, v') = f(v', v)^*$ for every $v, v' \in V$.
In this way a (unitary) dot product is characterized by an hermitian, positive definite sesquilinear form. The canonical unitary dot product in $\mathbb{C}^n$ is defined as $f(u, v) = u^{\mathrm{T}}v^*$.

If V is finite dimensional, then a sesquilinear form can be represented by a matrix (like in the real case):
If $f: V \times V \longrightarrow \mathbb{C}$ is a sesquilinear form, and $B = (b_1, b_2, \ldots, b_n)$ is a basis of V, then the matrix $\mathbf{A}_B(f) = (f(b_i, b_j))_{i,j}$ is the *representation matrix* of f with respect to basis B. The sesquilinear form can be written in matrix product form:

$$f(v, v') = v^{\mathrm{T}}\mathbf{A}_B(f)v', \qquad (5.213)$$

where v and v' are given with respect to basis B. A representation matrix is hermitian if and only if the sesquilinear form is hermitian.

5.4 Elementary Number Theory

Elementary number theory investigates divisibility properties of integers.

5.4.1 Divisibility

5.4.1.1 Divisibility and Elementary Divisibility Rules

1. Divisor

An integer $b \in \mathbb{Z}$ is *divisible* by an integer a without remainder iff* there is an integer q such that

$$qa = b \tag{5.214}$$

holds. Here a is a divisor of b in $\mathbb{Z}$, and q is the *complementary divisor* with respect to a; b is a *multiple* of a. For "a divides b" we write also $a|b$. For "a does not divide b" we can write $a \nmid b$. The divisibility relation (5.214) is a binary relation in $\mathbb{Z}$ (see 5.2.3, **2.**, p. 331). Analogously, divisibility is defined in the set of natural numbers.

2. Elementary Divisibility Rules

(DR1) For every $a \in \mathbb{Z}$ we have $1|a,\ a|a$ and $a|0$. (5.215)

(DR2) If $a|b,$ then $(-a)|b$ and $a|(-b)$. (5.216)

(DR3) $a|b$ and $b|a$ implies $a = b$ or $a = -b$. (5.217)

(DR4) $a|1$ implies $a = 1$ or $a = -1$. (5.218)

(DR5) $a|b$ and $b \neq 0$ imply $|a| \leq |b|$. (5.219)

(DR6) $a|b$ implies $a|zb$ for every $z \in \mathbb{Z}$. (5.220)

(DR7) $a|b$ implies $az|bz$ for every $z \in \mathbb{Z}$. (5.221)

(DR8) $az|bz$ and $z \neq 0$ implies $a|b$ for every $z \in \mathbb{Z}$. (5.222)

(DR9) $a|b$ and $b|c$ imply $a|c$. (5.223)

(DR10) $a|b$ and $c|d$ imply $ac|bd$. (5.224)

(DR11) $a|b$ and $a|c$ imply $a|(z_1b + z_2c)$ for arbitrary $z_1, z_2 \in \mathbb{Z}$. (5.225)

(DR12) $a|b$ and $a|(b + c)$ imply $a|c$. (5.226)

5.4.1.2 Prime Numbers

1. Definition and Properties of Prime Numbers

A positive integer p $(p > 1)$ is called a *prime number* iff 1 and p are its only divisors in the set $\mathbb{N}$ of positive integers. Positive integers which are not prime numbers are called *composite numbers*.
For every integer, the smallest positive divisor different from 1 is a prime number. There are infinitely many prime numbers.
A positive integer p $(p > 1)$ is a prime number iff for arbitrary positive integers a, b, $p|(ab)$ implies $p|a$ or $p|b$.

2. Sieve of Eratosthenes

By the method of the "*Sieve of Eratosthenes*", every prime number smaller than a given positive integer n can be determined:

a) Write down the list of all positive integers from 2 to n.

b) Underline 2 and delete every subsequent multiple of 2.

c) If p is the first non-deleted and non-underlined number, then underline p and delete every p-th number (beginning with $2p$ and counting the numbers of the original list).

d) Repeat step c) for every p $(p \leq \sqrt{n})$ and stop the algorithm.

*if and only if

Every underlined and non-deleted number is a prime number. In this way, all prime numbers $\leq n$ are obtained.
The prime numbers are called *prime elements* of the set of integers.

3. Prime Pairs

Prime numbers with a difference of 2 form *prime pairs* (twin primes).

■ $(3,5), (5,7), (11,13), (17,19), (29,31), (41,43), (59,61), (71,73), (101,103)$ are prime pairs.

4. Prime Triplets

Prime triplets consist of three prime numbers occuring among four consecutive odd numbers.

■ $(5,7,11), (7,11,13), (11,13,17), (13,17,19), (17,19,23), (37,41,43)$ are prime triplets.

5. Prime Quadruplets

If the first two and the last two of five consecutive odd numbers are prime pairs, then they are called a *prime quadruplet*.

■ $(5,7,11,13), (11,13,17,19), (101,103,107,109), (191,193,197,199)$ are prime quadruplets.

The conjecture that there exist infinitely many prime pairs, prime triplets, and prime quadruplets, is not proved still.

6. Mersenne Primes

If $2^k - 1,\ \ k \in \mathbb{N}$, is a prime number, then k is also a prime number. The numbers $2^p - 1$ (p prime) are called Mersenne numbers. A Mersenne prime is a Mersenne number $2^p - 1$ which is itself a prime number.

■ $2^p - 1$ is a prime number for the first ten values of p: 2, 3, 5, 7, 13,17, 19, 31, 61, 89, 107, etc.

Remark: Since a few years the largest known prime is always a Mersenne prime, e.g. $2^{43112609} - 1$ in 2008, $2^{57885161} - 1$ in 2013. In contrary to other natural numbers the numbers of the form $2^k - 1$ can be tested in a relatively simple way whether they are primes: Let $p > 2$ be a prime and a sequence of natural numbers is defined by $s_1 = 4$, $s_{i+1} := s_i^2 - 2$ $(i \geq 1)$. The number $2^p - 1$ is a prime if and only if the term of the sequence s_{p-1} is divisible by $2^p - 1$.
The prime test based on this statement is called Lucas-Lehmer test .

7. Fermat Primes

If a number $2^k+1,\ \ k \in \mathbb{N}$, is an odd prime number, then k is a power of 2. The numbers $2^k+1,\ \ k \in \mathbb{N}$, are called *Fermat numbers*. If a Fermat number is a prime number, then it is called a *Fermat prime*.

■ For $k = 0,1,2,3,4$ the corresponding Fermat numbers $3, 5, 17, 257, 65537$ are prime numbers. It is conjectured that there are no further Fermat primes.

8. Fundamental Theorem of Elementary Number Theory

Every positive integer $n > 1$ can be represented as a product of primes. This representation is unique except for the order of the factors. Therefore n is called to have exactly one *prime factorization*.

■ $360 = 2 \cdot 2 \cdot 2 \cdot 3 \cdot 3 \cdot 5 = 2^3 \cdot 3^2 \cdot 5.$

Remark: Analogously, the integers (except $-1, 0, 1$) can be represented as products of prime elements, unique apart from the order and the sign of the factors.

9. Canonical Prime Factorization

It is usual to arrange the factors of the prime factorization of a positive integer according to their size, and to combine equal factors to powers. If every non-occurring prime is assigned exponent 0, then every positive integer is uniquely determined by the sequence of the exponents of its prime factorization.

■ To $1\,533\,312 = 2^7 \cdot 3^2 \cdot 11^3$ belongs the sequence of exponents $(7,2,0,0,3,0,0,\ldots)$.

For a positive integer n, let $p_1, p_2, \ldots p_m$ be the pairwise distinct primes divisors of n, and let α_k denote the exponent of a prime number p_k in the prime factorization of n. Then

$$n = \prod_{k=1}^{m} p_k^{\alpha_k}, \tag{5.227a}$$

and this representation is called the *canonical prime factorization* of n. It is often denoted by

$$n = \prod_p p^{\nu_p(n)}, \tag{5.227b}$$

where the product applies to all prime numbers p, and where $\nu_p(n)$ is the multiplicity of p as a divisor of n. It always means a finite product because only finitely many of the exponents $\nu_p(n)$ differ from 0.

10. Positive Divisors

If a positive integer $n \geq 1$ is given by its canonical prime factorization (5.227a), then every positive divisor t of n can be written in the form

$$t = \prod_{k=1}^{m} p_k^{\tau_k} \quad \text{with } \tau_k \in \{0, 1, 2, \ldots, \alpha_k\} \text{ for } k = 1, 2, \ldots, m. \tag{5.228a}$$

The number $\tau(n)$ of all positive divisors of n is

$$\tau(n) = \prod_{k=1}^{m} (\alpha_k + 1). \tag{5.228b}$$

■ **A:** $\tau(5040) = \tau(2^4 \cdot 3^2 \cdot 5 \cdot 7) = (4+1)(2+1)(1+1)(1+1) = 60.$

■ **B:** $\tau(p_1 p_2 \cdots p_r) = 2^r$, if $p_1, p_2, \ldots, p_r$ are pairwise distinct prime numbers.

The product $P(n)$ of all positive divisors of n is given by

$$P(n) = n^{\frac{1}{2}\tau(n)}. \tag{5.228c}$$

■ **A:** $P(20) = 20^3 = 8000.$ ■ **B:** $P(p^3) = p^6$, if p is a prime number.

■ **C:** $P(pq) = p^2 q^2$, if p and q are different prime numbers.

The sum $\sigma(n)$ of all positive divisors of n is

$$\sigma(n) = \prod_{k=1}^{m} \frac{p_k^{\alpha_k+1} - 1}{p_k - 1}. \tag{5.228d}$$

■ **A:** $\sigma(120) = \sigma(2^3 \cdot 3 \cdot 5) = 15 \cdot 4 \cdot 6 = 360.$ ■ **B:** $\sigma(p) = p + 1$, if p is a prime number.

5.4.1.3 Criteria for Divisibility

1. Notation

Consider a positive integer given in decimal form:

$$n = (a_k a_{k-1} \cdots a_2 a_1 a_0)_{10} = a_k 10^k + a_{k-1} 10^{k-1} + \cdots + a_2 10^2 + a_1 10 + a_0. \tag{5.229a}$$

Then

$$Q_1(n) = a_0 + a_1 + a_2 + \cdots + a_k \tag{5.229b}$$

and

$$Q_1'(n) = a_0 - a_1 + a_2 - + \cdots + (-1)^k a_k \tag{5.229c}$$

are called the *sum of the digits* (*of first order*) and the *alternating sum of the digits* (*of first order*) of n, respectively. Furthermore,

$$Q_2(n) = (a_1 a_0)_{10} + (a_3 a_2)_{10} + (a_5 a_4)_{10} + \cdots \quad \text{and} \tag{5.229d}$$

$$Q_2'(n) = (a_1 a_0)_{10} - (a_3 a_2)_{10} + (a_5 a_4)_{10} - + \cdots \tag{5.229e}$$

are called the *sum of the digits and the alternating sum of the digits, respectively, of second order* and

$$Q_3(n) = (a_2 a_1 a_0)_{10} + (a_5 a_4 a_3)_{10} + (a_8 a_7 a_6)_{10} + \cdots \tag{5.229f}$$

and

$$Q_3'(n) = (a_2 a_1 a_0)_{10} - (a_5 a_4 a_3)_{10} + (a_8 a_7 a_6)_{10} - + \cdots \tag{5.229g}$$

are called the *sum of the digits and alternating sum of the digits, respectively, of third order* .

■ The number 123 456 789 has the following sum of the digits: $Q_1 = 9+8+7+6+5+4+3+2+1 = 45$, $Q_1' = 9-8+7-6+5-4+3-2+1 = 5$, $Q_2 = 89+67+45+23+1 = 225$, $Q_2' = 89-67+45-23+1 = 45$, $Q_3 = 789+456+123 = 1368$ and $Q_3' = 789-456+123 = 456$.

2. Criteria for Divisibility

There are the following criteria for divisibility:

DC-1: $3|n \Leftrightarrow 3|Q_1(n)$, (5.230a)

DC-2: $7|n \Leftrightarrow 7|Q_3'(n)$, (5.230b)

DC-3: $9|n \Leftrightarrow 9|Q_1(n)$, (5.230c)

DC-4: $11|n \Leftrightarrow 11|Q_1'(n)$, (5.230d)

DC-5: $13|n \Leftrightarrow 13|Q_3'(n)$ (5.230e)

DC-6: $37|n \Leftrightarrow 37|Q_3(n)$, (5.230f)

DC-7: $101|n \Leftrightarrow 101|Q_2'(n)$, (5.230g)

DC-8: $2|n \Leftrightarrow 2|a_0$, (5.230h)

DC-9: $5|n \Leftrightarrow 5|a_0$, (5.230i)

DC-10: $2^k|n \Leftrightarrow 2^k|(a_{k-1}a_{k-2}\cdots a_1a_0)_{10}$, (5.230j)

DC-11: $5^k|n \Leftrightarrow 5^k|(a_{k-1}a_{k-2}\cdots a_1a_0)_{10}$. (5.230k)

■ **A:** $a = 123\,456\,789$ is divisible by 9 since $Q_1(a) = 45$ and $9|45$, but it is not divisible by 7 since $Q_3'(a) = 456$ and $7\nmid 456$.

■ **B:** 91 619 is divisible by 11 since $Q_1'(91\,619) = 22$ and $11|22$.

■ **C:** 99 994 096 is divisible by 2^4 since $2^4|4\,096$.

5.4.1.4 Greatest Common Divisor and Least Common Multiple

1. Greatest Common Divisor

For integers $a_1, a_2, \ldots, a_n$, which are not all equal to zero, the largest number in the set of common divisors of $a_1, a_2, \ldots, a_n$ is called the *greatest common divisor* of $a_1, a_2, \ldots, a_n$, and it is denoted by $\gcd(a_1, a_2, \ldots, a_n)$. If $\gcd(a_1, a_2, \ldots, a_n) = 1$, then the numbers $a_1, a_2, \ldots, a_n$ are called *coprimes.*
To determine the greatest common divisor, it is sufficient to consider the positive common divisors. If the canonical prime factorizations

$$a_i = \prod_p p^{\nu_p(a_i)} \tag{5.231a}$$

of $a_1, a_2, \ldots, a_n$ are given, then

$$\gcd(a_1, a_2, \ldots, a_n) = \prod_p p^{\left\{\min\limits_i [\nu_p(a_i)]\right\}}. \tag{5.231b}$$

■ For the numbers $a_1 = 15\,400 = 2^3 \cdot 5^2 \cdot 7 \cdot 11$, $a_2 = 7\,875 = 3^2 \cdot 5^3 \cdot 7$, $a_3 = 3\,850 = 2 \cdot 5^2 \cdot 7 \cdot 11$, the greatest common divisor is $\gcd(a_1, a_2, a_3) = 5^2 \cdot 7 = 175$.

2. Euclidean Algorithm

The greatest common divisor of two integers a, b can be determined by the *Euclidean algorithm* without using their prime factorization. To do this, a sequence of divisions with remainder, according to the following scheme, is performed. For $a > b$ let $a_0 = a, a_1 = b$. Then:

$$\begin{aligned}
a_0 &= q_1a_1 + a_2, && 0 < a_2 < a_1,\\
a_1 &= q_2a_2 + a_3, && 0 < a_3 < a_2,\\
&\vdots \quad \vdots \quad \vdots\\
a_{n-2} &= q_{n-1}a_{n-1} + a_n, && 0 < a_n < a_{n-1},\\
a_{n-1} &= q_na_n.
\end{aligned} \tag{5.232a}$$

The division algorithm stops after a finite number of steps, since the sequence $a_2, a_3, \dots$ is a strictly monotone decreasing sequence of positive integers. The last remainder a_n, different from 0 is the greatest common divisor of a_0 and a_1.

■ gcd(38, 105) = 1, as can be seen by the help of the table to the right.

$$\begin{aligned} 105 &= 2 \cdot 38 + 29 \\ 38 &= 1 \cdot 29 + 9 \\ 29 &= 3 \cdot 9 + 2 \\ 9 &= 4 \cdot 2 + 1 \\ 2 &= 2 \cdot 1 \end{aligned}$$

By the recursion formula

$$\gcd(a_1, a_2, \dots, a_n) = \gcd(\gcd(a_1, a_2, \dots, a_{n-1}), a_n), \tag{5.232b}$$

the greatest common divisor of n positive integers with $n > 2$ can be determined by repeated use of the Euclidean algorithm.

■ $\gcd(150, 105, 56) = \gcd(\mathit{gcd}(150, 105), 56) = \gcd(15, 56) = 1.$

■ The Euclidean algorithm to determine the gcd (see also 1.1.1.4, **1.**, p. 3) of two numbers has especially many steps, if the numbers are adjacent numbers in the sequence of Fibonacci numbers (see 5.4.1.5, p. 375). The annexed calculation shows an example where all quotients are always equal to 1.

$$\begin{aligned} 55 &= 1 \cdot 34 + 21 \\ 34 &= 1 \cdot 21 + 13 \\ 21 &= 1 \cdot 13 + 8 \\ 13 &= 1 \cdot 8 + 5 \\ 8 &= 1 \cdot 5 + 3 \\ 5 &= 1 \cdot 3 + 2 \\ 3 &= 1 \cdot 2 + 1 \\ 2 &= 1 \cdot 1 + 1 \\ 1 &= 1 \cdot 1. \end{aligned}$$

3. Theorem for the Euclidean Algorithm

For two natural numbers a, b with $a > b > 0$, let $\lambda(a, b)$ denote the number of divisions with remainder in the Euclidean algorithm, and let $\kappa(b)$ denote the number of digits of b in the decimal system. Then

$$\lambda(a, b) \le 5 \cdot \kappa(b). \tag{5.233}$$

4. Greatest Common Divisor as a Linear Combination

It follows from the Euclidean algorithm that

$$\begin{aligned} a_2 &= a_0 - q_1 a_1 = c_0 a_0 + d_0 a_1, \\ a_3 &= a_1 - q_2 a_2 = c_1 a_0 + d_1 a_1, \\ &\vdots \quad \vdots \\ a_n &= a_{n-2} - q_{n-1} a_{n-1} = c_{n-2} a_0 + d_{n-2} a_1. \end{aligned} \tag{5.234a}$$

Here c_{n-2} and d_{n-2} are integers. Thus the $\gcd(a_0, a_1)$ can be represented as a linear combination of a_0 and a_1 with integer coefficients:

$$\gcd(a_0, a_1) = c_{n-2} a_0 + d_{n-2} a_1. \tag{5.234b}$$

Moreover $\gcd(a_1, a_2, \dots, a_n)$ can be represented as a linear combination of $a_1, a_2, \dots, a_n$, since:

$$\gcd(a_1, a_2, \dots, a_n) = \gcd(\gcd(a_1, a_2, \dots, a_{n-1}), a_n) = c \cdot \gcd(a_1, a_2, \dots, a_{n-1}) + d a_n. \tag{5.234c}$$

■ $\gcd(150, 105, 56) = \gcd(\gcd(150, 105), 56) = \gcd(15, 56) = 1$ with $15 = (-2) \cdot 150 + 3 \cdot 105$ and $1 = 15 \cdot 15 + (-4) \cdot 56$), thus $\gcd(150, 105, 56) = (-30) \cdot 150 + 45 \cdot 105 + (-4) \cdot 56$.

5. Least Common Multiple

For integers $a_1, a_2, \dots, a_n$, among which there is no zero, the smallest number in the set of positive common multiples of $a_1, a_2, \dots, a_n$ is called the *least common multiple* of $a_1, a_2, \dots, a_n$, and it is denoted by $\mathrm{lcm}(a_1, a_2, \dots, a_n)$.

If the canonical prime factorizations (5.231a) of $a_1, a_2, \dots, a_n$ are given, then:

$$\mathrm{lcm}(a_1, a_2, \dots, a_n) = \prod_p p^{\left\{ \max\limits_i [\nu_p(a_i)] \right\}}. \tag{5.235}$$

■ For the numbers $a_1 = 15\,400 = 2^3 \cdot 5^2 \cdot 7 \cdot 11$, $a_2 = 7\,875 = 3^2 \cdot 5^3 \cdot 7$, $a_3 = 3\,850 = 2 \cdot 5^2 \cdot 7 \cdot 11$ the least common multiple is $\mathrm{lcm}(a_1, a_2, a_3) = 2^3 \cdot 3^2 \cdot 5^3 \cdot 7 \cdot 11 = 693\,000$.

6. Relation between gcd and lcm

For arbitrary integers a, b:

$$|ab| = \gcd(a,b) \cdot \operatorname{lcm}(a,b). \tag{5.236}$$

Therefore, the $\operatorname{lcm}(a,b)$ can be determined with the help of the Euclidean algorithm without using the prime factorizations of a and b.

5.4.1.5 Fibonacci Numbers

1. Recursion Formula

The sequence

$$(F_n)_{n\in\mathbb{N}} \text{ with } F_1 = F_2 = 1 \text{ and } F_{n+2} = F_n + F_{n+1} \tag{5.237}$$

is called *Fibonacci sequence*. It starts with the elements $1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144, 233, 377, \ldots$

■ The consideration of this sequence goes back to the question posed by Fibonacci in 1202: How many pairs of descendants has a pair of rabbits at the end of a year, if every pair in every month produces a new pair, which beginning with the second month itself produces new descended pairs? The answer is $F_{14} = 377$.

2. Explicit Formula

Besides the recursive definition (5.237) there is an explicit formula for the Fibonacci numbers:

$$F_n = \frac{1}{\sqrt{5}}\left(\left[\frac{1+\sqrt{5}}{2}\right]^n - \left[\frac{1-\sqrt{5}}{2}\right]^n\right). \tag{5.238}$$

Some important properties of Fibonacci numbers are the followings. For $m, n \in \mathbb{N}$:

(1) $F_{m+n} = F_{m-1}F_n + F_mF_{n+1} \quad (m > 1).$ (5.239a)

(2) $F_m | F_{mn}.$ (5.239b)

(3) $\gcd(m,n) = d$ implies $\gcd(F_m, F_n) = F_d.$ (5.239c)

(4) $\gcd(F_n, F_{n+1}) = 1.$ (5.239d)

(5) $F_m | F_k$ holds iff $m|k$ holds. (5.239e)

(6) $\sum_{i=1}^{n} F_i^2 = F_nF_{n+1}.$ (5.239f)

(7) $\gcd(m,n) = 1$ implies $F_mF_n | F_{mn}.$ (5.239g)

(8) $\sum_{i=1}^{n} F_i = F_{n+2} - 1.$ (5.239h)

(9) $F_nF_{n+2} - F_{n+1}^2 = (-1)^{n+1}.$ (5.239i)

(10) $F_n^2 + F_{n+1}^2 = F_{2n+1}.$ (5.239j)

(11) $F_{n+2}^2 - F_n^2 = F_{2n+2}.$ (5.239k)

5.4.2 Linear Diophantine Equations

1. Diophantine Equations

An equation $f(x_1, x_2, \ldots, x_n) = b$ is called a *Diophantine equation* in n unknowns iff $f(x_1, x_2, \ldots, x_n)$ is a polynomial in $x_1, x_2, \ldots, x_n$ with coefficients in the set Z of integers, b is an integer constant and only integer solutions are of interest. The name “Diophantine” reminds of the Greek mathematician Diophantus, who lived around 250 AD.

In practice, Diophantine equations occur for instance, if relations between quantities are described. Until now, only general solutions of Diophantine equations of at most second degree with two variables are known. Solutions of Diophantine equations of higher degrees are only known in special cases.

2. Linear Diophantine Equations in n Unknowns

A *linear Diophantine equation* in n unknowns is an equation of the form

$$a_1x_1 + a_2x_2 + \cdots a_nx_n = b \quad (a_i \in \mathrm{Z},\ b \in \mathrm{Z}), \tag{5.240}$$

where only integer solutions are searched for. A solution method is described in the following.

3. Conditions of Solvability

If not all the coefficients a_i are equal to zero, then the Diophantine equation (5.240) is solvable iff $\gcd(a_1, a_2, \ldots, a_n)$ is a divisor of b.

■ $114x + 315y = 3$ is solvable, since $\gcd(114, 315) = 3$.

If a linear Diophantine equation in n unknowns ($n > 1$) has a solution and Z is the domain of variables, then the equation has infinitely many solutions. Then in the set of solutions there are $n-1$ free variables. For subsets of Z, this statement is not true.

4. Solution Method for $n = 2$

Let

$$a_1x_1 + a_2x_2 = b \quad (a_1, a_2) \neq (0, 0) \tag{5.241a}$$

be a solvable Diophantine equation, i.e., $\gcd(a_1, a_2)|b$. To find a special solution of the equation, the equation is divided by $\gcd(a_1, a_2)$ and one obtains $a_1'x_1' + a_2'x_2' = b'$ with $\gcd(a_1', a_2') = 1$.
As described in 5.4.1, **4.**, p. 374, $\gcd(a_1', a_2')$ is determined to obtain finally a linear combination of a_1' and a_2': $a_1'c_1' + a_2'c_2' = 1$.
Substitution in the given equation demonstrates that the ordered pair $(c_1'b', c_2'b')$ of integers is a solution of the given Diophantine equation.

■ $114x+315y = 6$. The equation is divided by 3, since $3 = \gcd(114, 315)$. That implies $38x+105y = 2$ and $38 \cdot 47 + 105 \cdot (-17) = 1$ (see 5.4.1, **4.**, p. 374). The ordered pair $(47 \cdot 2, (-17) \cdot 2) = (94, -34)$ is a special solution of the equation $114x + 315y = 6$.

The family of solutions of (5.241a) can be obtained as follows: If (x_1^0, x_2^0) is an arbitrary special solution, which could also be obtained by trial and error, then

$$\{(x_1^0 + t \cdot a_2', x_2^0 - t \cdot a_1')|t \in \mathrm{Z}\} \tag{5.241b}$$

is the set of all solutions.

■ The set of solutions of the equation $114x + 315y = 6$ is $\{(94 + 315t, -34 - 114t)|t \in \mathrm{Z}\}$.

5. Reduction Method for $n > 2$

Suppose a solvable Diophantine equation

$$a_1x_1 + a_2x_2 + \cdots + a_nx_n = b \tag{5.242a}$$

with $(a_1, a_2, \ldots, a_n) \neq (0, 0, \ldots, 0)$ and $\gcd(a_1, a_2, \ldots, a_n) = 1$ is given. If $\gcd(a_1, a_2, \ldots, a_n) \neq 1$, then the equation should be divided by $\gcd(a_1, a_2, \ldots, a_n)$. After the transformation

$$a_1x_1 + a_2x_2 + \cdots + a_{n-1}x_{n-1} = b - a_nx_n \tag{5.242b}$$

x_n is considered as an integer constant and a linear Diophantine equation in $n-1$ unknowns is obtained, and it is solvable iff $\gcd(a_1, a_2, \ldots, a_{n-1})$ is a divisor of $b - a_nx_n$.
The condition

$$\gcd(a_1, a_2, \ldots, a_{n-1})|b - a_nx_n \tag{5.242c}$$

is satisfied iff there are integers $\underline{c}, \underline{c}_n$ such that:

$$\gcd(a_1, a_2, \ldots, a_{n-1}) \cdot \underline{c} + a_n\underline{c}_n = b. \tag{5.242d}$$

This is a linear Diophantine equation in two unknowns, and it can be solved as shown in 5.4.2,**4.**, p. 376. If its solution is determined, then it remains to solve a Diophantine equation in only $n - 1$ unknowns. This procedure can be continued until a Diophantine equation in two unknowns is obtained, which can be solved with the method given in 5.4.2, **4.**, p. 376.
Finally, the solution of the given equation is constructed from the set of solutions obtained in this way.

■ Solve the Diophantine equation

$$2x + 4y + 3z = 3. \tag{5.243a}$$

This is solvable since gcd(2, 4, 3) is a divisor of 3.
The Diophantine equation

$$2x + 4y = 3 - 3z \tag{5.243b}$$

in the unknowns x, y is solvable iff gcd(2, 4) is a divisor of $3 - 3z$. The corresponding Diophantine equation $2z' + 3z = 3$ has the set of solutions $\{(-3 + 3t, 3 - 2t) | t \in \mathbb{Z}\}$. This implies, $z = 3 - 2t$, and now the set of solutions of the solvable Diophantine equation $2x + 4y = 3 - 3(3 - 2t)$ or

$$x + 2y = -3 + 3t \tag{5.243c}$$

is sought for every $t \in \mathbb{Z}$.
The equation (5.243c) is solvable since $\gcd(1, 2) = 1 | (-3 + 3t)$. Now $1 \cdot (-1) + 2 \cdot 1 = 1$ and $1 \cdot (3 - 3t) + 2 \cdot (-3 + 3t) = -3 + 3t$. The set of solution is $\{((3 - 3t) + 2s, (-3 + 3t) - s) | s \in \mathbb{Z}\}$. That implies $x = (3 - 3t) + 2s, y = (-3 + 3t) - s$, and $\{(3 - 3t + 2s, -3 + 3t - s, 3 - 2t) | s, t \in \mathbb{Z}\}$ so obtained is the set of solutions of (5.243a).

5.4.3 Congruences and Residue Classes

1. Congruences

Let m be a positive integer m, $m > 1$. If two integers a and b have the same remainder, when divided by m, then a and b are called *congruent modulo* m, denoted by $a \equiv b \bmod m$ or $a \equiv b(m)$.

■ $3 \equiv 13 \bmod 5, \quad 38 \equiv 13 \bmod 5, \quad 3 \equiv -2 \bmod 5.$

Remark: Obviously, $a \equiv b \bmod m$ holds iff m is a divisor of the difference $a - b$. Congruence modulo m is an equivalence relation (see 5.2.4, **1.**, p. 334) in the set of integers. Note the following properties:

$$a \equiv a \bmod m \text{ for every } a \in \mathbb{Z}, \tag{5.244a}$$

$$a \equiv b \bmod m \Rightarrow b \equiv a \bmod m, \tag{5.244b}$$

$$a \equiv b \bmod m \wedge b \equiv c \bmod m \Rightarrow a \equiv c \bmod m. \tag{5.244c}$$

2. Calculating Rules

$$a \equiv b \bmod m \wedge c \equiv d \bmod m \Rightarrow a + c \equiv b + d \bmod m, \tag{5.245a}$$

$$a \equiv b \bmod m \wedge c \equiv d \bmod m \Rightarrow a \cdot c \equiv b \cdot d \bmod m, \tag{5.245b}$$

$$a \cdot c \equiv b \cdot c \bmod m \wedge \gcd(c, m) = 1 \Rightarrow a \equiv b \bmod m, \tag{5.245c}$$

$$a \cdot c \equiv b \cdot c \bmod m \wedge c \neq 0 \Rightarrow a \equiv b \bmod \frac{m}{\gcd(c, m)}. \tag{5.245d}$$

3. Residue Classes, Residue Class Ring

Since congruence modulo m is an equivalence relation in $\mathbb{Z}$, this relation induces a partition of $\mathbb{Z}$ into *residue classes modulo m*:

$$[a]_m = \{x | x \in \mathbb{Z} \wedge x \equiv a \bmod m\}. \tag{5.246}$$

The residue class " a modulo m " consists of all integers having equal remainder if divided by m. Now $[a]_m = [b]_m$ iff $a \equiv b \bmod m$.
There are exactly m residue classes modulo m, and normally they are represented by their smallest non-negative representatives:

$$[0]_m, [1]_m, \ldots, [m-1]_m. \tag{5.247}$$

In the set $\mathbb{Z}_m$ of residue classes modulo m, *residue class addition* and *residue class multiplication* are defined by

$$[a]_m \oplus [b]_m := [a + b]_m, \tag{5.248}$$

$$[a]_m \odot [b]_m := [a \cdot b]_m. \tag{5.249}$$

These residue class operations are independent of the chosen representatives, i.e.,

$$[a]_m = [a']_m \text{ and } [b]_m = [b']_m \text{ imply}$$
$$[a]_m \oplus [b]_m = [a']_m \oplus [b']_m \text{ and } [a]_m \odot [b]_m = [a']_m \odot [b']_m. \quad (5.250)$$

The residue classes modulo m form a ring with unit element, with respect to residue class addition and residue class multiplication (see 5.4.3, **1.**, p. 377), the *residue class ring modulo m*. If p is a prime number, then the residue class ring modulo p is a field (see 5.3.7, **2.**, p. 361).

4. Residue Classes Relatively Prime to *m*

A residue class $[a]_m$ with $\gcd(a, m) = 1$ is called a *residue class relatively prime to m*. If p is a prime number, then all residue classes different from $[0]_p$ are residue classes relatively prime to p.
The residue classes relatively prime to m form an Abelian group (5.3.3.1,**1.**, p. 336) with respect to residue class multiplication, the so-called *group of residue classes relatively prime to m*. The order of this group is $\varphi(m)$, where φ is the *Euler function* (see 5.4.4, **1.**, p. 381).

■ **A:** $[1]_8, [3]_8, [5]_8, [7]_8$ are residue classes relatively prime to 8.

■ **B:** $[1]_5, [2]_5, [3]_5, [4]_5$ are residue classes relatively prime to 5.

■ **C:** $\varphi(8) = \varphi(5) = 4$ is valid.

5. Primitive Residue Classes

A residue class $[a]_m$ relatively prime to m is called a *primitive residue class* if it has order $\varphi(m)$ in the group of residue classes relatively prime to m.

■ **A:** $[2]_5$ is a primitive residue class modulo 5, since $([2]_5)^2 = [4]_5,\ ([2]_5)^3 = [3]_5,\ ([2]_5)^4 = [1]_5$.

■ **B:** There is no primitive residue class modulo 8, since $[1]_8$ has order 1, and $[3]_8, [5]_8, [7]_8$ have order 2 in the group of residue classes relatively prime to m.

Remark: There is a primitive residue class modulo m, iff $m = 2, m = 4, m = p^k$ or $m = 2p^k$, where p is an odd prime number and k is a positive integer.
If there is a primitive residue class modulo m, then the group of residue classes relatively prime to m forms a cyclic group.

6. Linear Congruences

1. Definition If a, b and $m > 0$ are integers, then

$$ax \equiv b(m) \quad (5.251)$$

is called a *linear congruence* (*in the unknown* x).

2. Solutions An integer x^* satisfying $ax^* \equiv b(m)$ is a solution of this congruence. Every integer, which is congruent to x^* modulo m, is also a solution. In finding all solutions of (5.251) it is sufficient to find the integers pairwise incongruent modulo m which satisfy the congruence.
The congruence (5.251) is solvable iff $\gcd(a, m)$ is a divisor of b. In this case, the number of solutions modulo m is equal to $\gcd(a, m)$.
In particular, if $\gcd(a, m) = 1$ holds, the congruence modulo m has a unique solution.

3. Solution Method There are different solution methods for linear congruences. It is possible to transform the congruence $ax \equiv b(m)$ into the Diophantine equation $ax + my = b$, and to determine a special solution (x^0, y^0) of the Diophantine equation $a'x + m'y = b'$ with $a' = a/\gcd(a, m), m' = m/\gcd(a, m), b' = b/\gcd(a, m)$ (see 5.4.2, **1.**, p. 375).
The congruence $a'x \equiv b'(m')$ has a unique solution since $\gcd(a', m') = 1$ modulo m', and

$$x \equiv x^0(m'). \quad (5.252a)$$

The congruence $ax \equiv b(m)$ has exactly $\gcd(a, m)$ solutions modulo m:

$$x^0, x^0 + m, x^0 + 2m, \ldots, x^0 + (\gcd(a, m) - 1)m. \quad (5.252b)$$

■ $114x \equiv 6 \bmod 315$ is solvable, since $\gcd(114, 315)$ is a divisor of 6; there are three solutions modulo 315.
$38x \equiv 2 \bmod 105$ has a unique solution: $x \equiv 94 \bmod 105$ (see 5.4.2, **4.**, p. 376). 94, 199, and 304 are the solutions of $114x \equiv 3 \bmod 315$.

7. Simultaneous Linear Congruences

If finitely many congruences

$$x \equiv b_1(m_1), x \equiv b_2(m_2), \ldots, x \equiv b_t(m_t) \quad (5.253)$$

are given, then (5.253) is called a *system of simultaneous linear congruences.* A result on the set of solutions is the *Chinese remainder theorem*: Consider a given system $x \equiv b_1(m_1), x \equiv b_2(m_2), \ldots, x \equiv b_t(m_t)$, where $m_1, m_2, \ldots, m_t$ are pairwise coprime numbers. If

$$m = m_1 \cdot m_2 \cdots m_t, a_1 = \frac{m}{m_1}, a_2 = \frac{m}{m_2}, \ldots, a_t = \frac{m}{m_t} \quad (5.254a)$$

and x_j is chosen such that $a_j x_j \equiv b_j(m_j)$ for $j = 1, 2, \ldots, t$, then

$$x' = a_1 x_1 + a_2 x_2 + \cdots + a_t x_t \quad (5.254b)$$

is a solution of the system. The system has a unique solution modulo m, i.e., if x' is a solution, then x'' is a solution, too, iff $x'' \equiv x'(m)$.

■ Solve the system $x \equiv 1\,(2)$, $x \equiv 2\,(3)$, $x \equiv 4\,(5)$, where $2, 3, 5$ are pairwise coprime numbers. Then $m = 30, a_1 = 15, a_2 = 10, a_3 = 6$. The congruences $15x_1 \equiv 1\,(2)$, $10x_2 \equiv 2\,(3)$, $6x_3 \equiv 4\,(5)$ have the special solutions $x_1 = 1$, $x_2 = 2$, $x_3 = 4$. The given system has a unique solution modulo m: $x \equiv 15 \cdot 1 + 10 \cdot 2 + 6 \cdot 4\,(30)$, i.e., $x \equiv 29\,(30)$.

Remark: Systems of simultaneous linear congruences can be used to reduce the problem of solving non-linear congruences modulo m to the problem of solving congruences modulo prime number powers (see 5.4.3, **9.**, p. 380).

8. Quadratic Congruences

1. Quadratic Residues Modulo m One can solve every congruence $ax^2 + bx + c \equiv 0(m)$ if one can solve every congruence $x^2 \equiv a(m)$:

$$ax^2 + bx + c \equiv 0(m) \Leftrightarrow (2ax + b)^2 \equiv b^2 - 4ac(m). \quad (5.255)$$

First quadratic residues modulo m are considered: Let $m \in \mathbb{N}, m > 1$ and $a \in \mathbb{Z}$, $\gcd(a, m) = 1$. The number a is called a *quadratic residue modulo m* iff there is an $x \in \mathbb{Z}$ with $x^2 \equiv a(m)$.

If the canonical prime factorization of m is given, i.e.,

$$m = \prod_{i=1}^{\infty} p_i^{\alpha_i}, \quad (5.256)$$

then r is a quadratic residue modulo m iff r is a quadratic residue modulo $p_i^{\alpha_i}$ for $i = 1, 2, 3, \ldots$.

If a is a quadratic residue modulo a prime number p, then this is denoted by $\left(\frac{a}{p}\right) = 1$; if a is a quadratic non-residue modulo p, then it is denoted by $\left(\frac{a}{p}\right) = -1$ (Legendre symbol).

■ The numbers $1, 4, 7$ are quadratic residues modulo 9.

2. Properties of Quadratic Congruences

(E1) $p \nmid ab$ and $a \equiv b(p)$ imply $\left(\frac{a}{p}\right) = \left(\frac{b}{p}\right)$. (5.257a)

(E2) $\left(\frac{1}{p}\right) = 1$. (5.257b)

(E3) $\left(\frac{-1}{p}\right) = (-1)^{\frac{p-1}{2}}$. (5.257c)

(E4) $\left(\frac{ab}{p}\right) = \left(\frac{a}{p}\right) \cdot \left(\frac{b}{p}\right)$ in particular $\left(\frac{ab^2}{p}\right) = \left(\frac{a}{p}\right)$. (5.257d)

(E5) $\left(\frac{2}{p}\right) = (-1)^{\frac{p^2-1}{8}}$. (5.257e)

(E6) Quadratic reciprocity law: If p and q are distinct odd prime numbers,

then $\left(\frac{p}{q}\right) \cdot \left(\frac{q}{p}\right) = (-1)^{\frac{p-1}{2}\frac{q-1}{2}}$. (5.257f)

■ $\left(\frac{65}{307}\right) = \left(\frac{5}{307}\right) \cdot \left(\frac{13}{307}\right) = \left(\frac{307}{5}\right) \cdot \left(\frac{307}{13}\right) = \left(\frac{2}{5}\right) \cdot \left(\frac{8}{13}\right) = (-1)^{\frac{5^2-1}{8}} \left(\frac{2^3}{13}\right) = -\left(\frac{2}{13}\right) = -(-1)^{\frac{13^2-1}{8}} = 1$.

In General: A congruence $x^2 \equiv a(2^\alpha)$, $\gcd(a,2) = 1$, is solvable iff $a \equiv 1(4)$ for $\alpha = 2$ and $a \equiv 1(8)$ for $\alpha \geq 3$. If these conditions are satisfied, then modulo 2^α there is one solution for $\alpha = 1$, there are two solutions for $\alpha = 2$ and four solutions for $\alpha \geq 3$.

A necessary condition for solvability of congruences of the general form

$$x^2 \equiv a(m),\ m = 2^\alpha p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_t^{\alpha_t}, \quad \gcd(a,m) = 1, \tag{5.258a}$$

is the solvability of the congruences

$$a \equiv 1(4) \text{ for } \alpha = 2, \quad a \equiv 1(8) \text{ for } \alpha \geq 3, \quad \left(\frac{a}{p_1}\right) = 1,\ \left(\frac{a}{p_2}\right) = 1,\ \ldots,\ \left(\frac{a}{p_t}\right) = 1. \tag{5.258b}$$

If all these conditions are satisfied, then the number of solutions is equal to 2^t for $\alpha = 0$ and $\alpha = 1$, equal to 2^{t+1} for $\alpha = 2$ and equal to 2^{t+2} for $\alpha \geq 3$.

9. Polynomial Congruences

If $m_1, m_2, \ldots, m_t$ are pairwise coprime numbers, then the congruence

$$f(x) \equiv a_n x^n + a_{n-1} x^{n-1} + \cdots + a_0 \equiv 0(m_1 m_2 \cdots m_t) \tag{5.259a}$$

is equivalent to the system

$$f(x) \equiv 0(m_1),\ f(x) \equiv 0(m_2),\ \ldots,\ f(x) \equiv 0(m_t). \tag{5.259b}$$

If k_j is the number of solutions of $f(x) \equiv 0(m_j)$ for $j = 1, 2, \ldots, t$, then $k_1 k_2 \cdots k_t$ is the number of solutions of $f(x) \equiv 0(m_1 m_2 \cdots m_t)$. This means that the solution of the congruence

$$f(x) \equiv 0\ (p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_t^{\alpha_t}), \tag{5.259c}$$

where $p_1, p_2, \ldots, p_t$ are primes, can be reduced to the solution of congruences $f(x) \equiv 0(p^\alpha)$. Moreover, these congruences can be reduced to congruences $f(x) \equiv 0(p)$ modulo prime numbers in the following way:

a) A solution of $f(x) \equiv 0(p^\alpha)$ is a solution of $f(x) \equiv 0(p)$, too.

b) A solution $x \equiv x_1(p)$ of $f(x) \equiv 0(p)$ defines a unique solution modulo p^α iff $f'(x_1)$ is not divisible by p:

Suppose $f(x_1) \equiv 0(p)$. Let $x = x_1 + pt_1$ and determine the unique solution t_1' of the linear congruence

$$\frac{f(x_1)}{p} + f'(x_1)t_1 \equiv 0(p). \tag{5.260a}$$

Substitute $t_1 = t_1' + pt_2$ into $x = x_1 + pt_1$, then $x = x_2 + p^2 t_2$ is obtained. Now, the solution t_2' of the linear congruence

$$\frac{f(x_2)}{p^2} + f'(x_2)t_2 \equiv 0(p) \tag{5.260b}$$

has to be determined modulo p^2. By substitution of $t_2 = t_2' + pt_3$ into $x = x_2 + p^2 t_2$ the result $x = x_3 + p^3 t_3$ is obtained. Continuing this process yields the solution of the congruence $f(x) \equiv 0\,(p^\alpha)$.

■ Solve the congruence $f(x) = x^4 + 7x + 4 \equiv 0\,(27)$. $f(x) = x^4 + 7x + 4 \equiv 0\,(3)$ implies $x \equiv 1\,(3)$, i.e., $x = 1 + 3t_1$. Because of $f'(x) = 4x^3 + 7$ and $3 \not| f'(1)$ now the solution of the congruence $f(1)/3 + f'(1) \cdot t_1 \equiv 4 + 11t_1 \equiv 0\,(3)$ is searched for: $t_1 \equiv 1\,(3)$, i.e., $t_1 = 1 + 3t_2$ and $x = 4 + 9t_2$. Then consider $f(4)/9 + f'(4) \cdot t_2 \equiv 0\,(3)$ and the solution $t_2 \equiv 2\,(3)$ is obtained, i.e., $t_2 = 2 + 3t_3$ and $x = 22 + 27t_3$. Therefore, 22 is the solution of $x^4 + 7x + 4 \equiv 0\,(27)$, uniquely determined modulo 27.

5.4.4 Theorems of Fermat, Euler, and Wilson

1. Euler Function

For every positive integer m with $m > 0$ one can determine the number of coprimes x with respect to m for $1 \le x \le m$. The corresponding function φ is called the Euler function. The value of the function $\varphi(m)$ is the number of residue classes relatively prime to m (s. 5.4.3, **4.**, p. 378).
For instance, $\varphi(1) = 1,\ \varphi(2) = 1,\ \varphi(3) = 2,\ \varphi(4) = 2,\ \varphi(5) = 4,\ \varphi(6) = 2,\ \varphi(7) = 6,\ \varphi(8) = 4$, etc. In general, $\varphi(p) = p - 1$ holds for every prime number p and $\varphi(p^\alpha) = p^\alpha - p^{\alpha-1}$ for every prime number power p^α. If m is an arbitrary positive integer, then $\varphi(m)$ can be determined in the following way:

$$\varphi(m) = m \prod_{p|m} \left(1 - \frac{1}{p}\right), \tag{5.261a}$$

where the product applies to all prime divisors p of m.

■ $\varphi(360) = \varphi(2^3 \cdot 3^2 \cdot 5) = 360 \cdot (1 - \frac{1}{2}) \cdot (1 - \frac{1}{3}) \cdot (1 - \frac{1}{5}) = 96$.

Furthermore

$$\sum_{d|m} \varphi(d) = m \tag{5.261b}$$

is valid. If $\gcd(m, n) = 1$ holds, then we get $\varphi(mn) = \varphi(m)\varphi(n)$.

■ $\varphi(360) = \varphi(2^3 \cdot 3^2 \cdot 5) = \varphi(2^3) \cdot \varphi(3^2) \cdot \varphi(5) = 4 \cdot 6 \cdot 4 = 96$.

2. Fermat-Euler Theorem

The *Fermat-Euler theorem* is one of the most important theorems of elementary number theory. If a and m are coprime positive numbers, then

$$a^{\varphi(m)} \equiv 1(m). \tag{5.262}$$

■ Determine the last three digits of 9^{9^9} in decimal notation. This means, determine x with $x \equiv 9^{9^9}\,(1000)$ and $0 \le x \le 999$. Now $\varphi(1000) = 400$, and according to Fermats theorem $9^{400} \equiv 1\,(1000)$. Furthermore $9^9 = (80 + 1)^4 \cdot 9 \equiv \left(\binom{4}{0} 80^0 \cdot 1^4 + \binom{4}{1} 80^1 \cdot 1^3\right) \cdot 9 = (1 + 4 \cdot 80) \cdot 9 \equiv -79 \cdot 9 \equiv 89\,(400)$. From that it follows that $9^{9^9} \equiv 9^{89} = (10 - 1)^{89} \equiv \binom{89}{0} 10^0 \cdot (-1)^{89} + \binom{89}{1} 10^1 \cdot (-1)^{88} + \binom{89}{2} 10^2 \cdot (-1)^{87} = -1 + 89 \cdot 10 - 3916 \cdot 100 \equiv -1 - 110 + 400 = 289(1000)$. The decimal notation of 9^{9^9} ends with the digits 289.

Remark: The theorem above for $m = p$, i.e., $\varphi(p) = p - 1$ was proved by Fermat; the general form was proved by Euler. This theorem forms the basis for encoding schemes (see 5.4.6). It contains a necessary criterion for the prime number property of a positive integer: If p is a prime, then $a^{p-1} \equiv 1(p)$ holds for every integer a with $p \not| a$.

3. Wilson's Theorem

There is a further prime number criterion, called the Wilson theorem:
Every prime number p satisfies $(p - 1)! \equiv -1(p)$.
The inverse proposition is also true; and therefore:

The number p is a prime number iff $(p-1)! \equiv -1(p)$.

5.4.5 Prime Number Tests

In the followings two stochastic prime tests will be presented which are useful at large numbers to test the prime property with a sufficiently small probability of mistakes. With these tests it is possible to show that a number is not a prime, without knowing its prime factors.

1. Fermat-Prime Number Test

Let n be an odd natural number and a an integer such that $\gcd(a,n) = 1$ and $a^{n-1} \equiv 1 \pmod{n}$.. Then n is called a *pseudoprime* to base a.

■ **A:** 341 is a pseudo prime to basis 2; 341 is not a pseudo prime to basis 3.

Test: Let an odd natural number $n > 1$ be given. Choose $a \in \mathbb{Z}_n \setminus \{0\}$.

• If the gcd $(a,n) > 1$, then n is not prime.

• If the gcd $(a,n) = 1$ and $\left\{ \begin{array}{ll} a^{n-1} \equiv 1 & (\text{mod } n) \\ a^{n-1} \not\equiv 1 & (\text{mod } n) \end{array} \right\}$, then $n \left\{ \begin{array}{c} \text{did pass} \\ \text{did not pass} \end{array} \right\}$ the test to base a. If n did not pass the test, then n is not a prime. If n did pass the test, then it may be a prime, but more tests are needed with other base, i.e. tests with further values of a.

■ **B:** $n = 15$: The test with $a = 4$ gives $4^{14} \equiv 1 \pmod{15}$. The test with $a = 7$ gives $7^{14} \equiv 4 \not\equiv 1 \pmod{15}$. Hence 15 is not a prime.

■ **C:** $n = 561$: The test with arbitrary $a \in \mathbb{Z}_{561} \setminus \{0\}$ with $\gcd(a, 561) = 1$ results in $a^{560} \equiv 1 \pmod{561}$. But $561 = 3 \cdot 11 \cdot 17$ is not a prime.

Remark: A composite number n for which $a^{n-1} \equiv 1 \pmod{n}$ for all $a \in \mathbb{Z}_n \setminus \{0\}$ with $\gcd(a,n) = 1$ is called a Carmichael number.

If n is not a prime and not a Carmichael number, then one can show that the level of error of the first kind to get a false result using k numbers with $\gcd(a,n)$ =1 is at most $1/2^k$. At least for the half of the numbers in $\mathbb{Z}_n \setminus \{0\}$ with $\gcd(a,n)$ =1 the relation $a^{n-1} \not\equiv 1 \pmod{n}$ holds.

2. Rabin-Miller Prim Number Test

The Rabin-Miller primality test is based on the following statement $(*)$:
Let $n > 2$ be a prime, $n - 1 = 2^t u$ (u is odd), g.c.d(a,n) =1. Then:

$$a^u \equiv 1 \pmod{n} \text{ or } a^{2^j u} \equiv -1 \pmod{n} \text{ for some } j \in \{0, 1, \ldots, t-1\}. \qquad (*)$$

Every odd natural number $n > 1$ can be tested about prime property in the following way:
Test: Choose $a \in \mathbb{Z}_n \setminus \{0\}$ and find the representation $n - 1 = 2^t u$ (u is odd).
• If g.c.d$(a,n) > 1$, then n is not a prime.

• If g.c.d(a,n) =1, then the sequence $a^u \pmod{n}$, $a^{2u} (\text{mod} n)$, ..., $a^{2^{t-1}u} \pmod{n}$ is calculated until a value is found which satisfies $(*)$. These elements are calculated by repeated squaring mod n. If there is no such value, then n is not a prime. Otherwise n did pass the test to basis a.

■ **A**: $n = 561$, and should be tested by different values of a:

$$n - 1 = 2^4 \cdot 35, \quad a = 2: \begin{array}{rl} 2^{35} \equiv 263 \not\equiv \pm 1 & (\text{mod } 561), \\ 2^{70} \equiv 166 \not\equiv -1 & (\text{mod } 561), \\ 2^{140} \equiv 67 \not\equiv -1 & (\text{mod } 561), \\ 2^{280} \equiv 421 \not\equiv -1 & (\text{mod } 561). \end{array} \quad \text{561 is not a prime.}$$

If choosing k different values randomly and independently and n passes the test to basis a for each, then the error rate of the first kind that n is not a prime is $\leq 1/4^k$. In the practice $k = 25$ is chosen.

■ **B:** There is only one number $\leq 2,5 \cdot 10^{10}$ such that it passes the test to basis $a = 2, 3, 5, 7$ and it is not a prime.

3. AKS Prime Number Test

The AKS primality test is based on a polynomial algorithm to determine whether a number is prime or composite. Published by Agrawal, Kayal, and Saxena, in 2002, meanwhile it is evident that the prime property can be tested efficiently for any natural number.

The test is based on the following statements:
If $n > 1$ is a natural number and r is a prime satisfying the assumtions
• n is not divisible by primes $\leq r$,
• $r^i \not\equiv 1 \pmod{n}$ for $i = 1, 2, \ldots, \lfloor(\log_2 n)^2\rfloor^*$,
• $(x+a)^n \equiv x^n + a \pmod{x^r - 1, n}$ for every $1 \leq a \leq \sqrt{r}\log n$,
Then n is a power of a prime.

Let $n > 1$ be an odd natural number whose prime characteristic is to be tested, and $m := \lfloor(\log_2 n)^5\rfloor$. If $n < 5690034$, then it is tested by comparing it to a list of known prime numbers whether n is a prime. For $n > 5690034$ holds $n > m$:

Test:
• Check, whether n can be divided by a natural number from the interval $[3, m]$. If yes, then n is not a prime.
• Otherwise take a prime $r < m$, such that $r^i \not\equiv 1(\bmod n)$ for $i = 1, 2, \ldots, \lfloor(\log_2 n)^2\rfloor$. (It can be proven, that such a prime r exists.)
• Check, whether the congruence $(x+a)^n = x^n + a(\bmod x^r - 1, n)$ for $a = 1, 2, \sqrt{r}\lfloor(\log_2 n)\rfloor$ holds. If not, then n is not a prime. If yes, then n is a power of a prime. In this case it is to be tested, whether natural numbers q and $k > 1$ exist, for which $n = q^k$. If not, then n is a prime.

Different to the known and efficient stochastic algorithms, the result of the test can be trusted without even a negligible small error probability of mistakes. However in cryptography the Rabin-Miller test is preferred.

5.4.6 Codes

5.4.6.1 Control Digits

In the information theory methods are provided to recognize and to correct errors in data combinations. Some of the simplest methods are represented in the form of the following control digits.

1. International Standard Book Number ISBN-10

A simple application of the congruence of numbers is the use of control digits with the International Standard Book Number ISBN. A combination of 10 digits of the form

$$\text{ISBN } a - bcd - efghi - p. \tag{5.263a}$$

is assigned to a book. The digits have the following meaning: a is the group number (for example, $a = 3$ tells us that the book originates from Austria, Germany, or Switzerland), bcd is the publisher's number, and $efghi$ is the title number of the book by this publisher. A control digit p will be added to detect erroneous book orders and thus help reduce expenses. The control digit p is the smallest non-negative digit that fulfils the following congruence:

$$10a + 9b + 8c + 7d + 6e + 5f + 4g + 3h + 2i + p \equiv 0(11). \tag{5.263b}$$

If the control digit p is 10, a unary symbol such as X is used (see also 5.4.6, **3.**, p. 384). A presented ISBN can now be checked for a match of the control digit contained in the ISBN and the control digit determined from all the other digits. In case of no match an error is certain. The ISBN control digit method permits the detection of the following errors:

1. Single digit error and

2. interchange of two digits.

Statistical investigations showed that by this method more than 90% of all actual errors can be detected. All other observed error types have a relative frequency of less than 1%. In the majority of the cases

---

* $\lfloor x \rfloor$ is symbol for "greatest integer $\leq x$".

the described method will detect the interchange of two digits or the interchange of two complete digit blocks.

2. Central Codes for Drugs and Medicines

In pharmacy, a similar numerical system with control digits is employed for identifying medicaments. In Germany, each medicament is assigned a seven digit control code:

$$abcdefp. \tag{5.264a}$$

The last digit is the control digit p. It is the smallest, non-negative number that fulfils the congruence

$$2a + 3b + 4c + 5d + 6e + 7f \equiv p(11). \tag{5.264b}$$

Here too, the single digit error or the interchange of two digits can always be detected.

3. Account Numbers

Banks and saving banks use a uniform account number system with a maximum of 10 digits (depending on the business volume). The first (at most four) digits serve the classification of the account. The remaining six digits represent the actual account number including a control digit in the last position. The individual banks and saving banks tend to apply different control digit methods, for example:

a) The digits are multiplied alternately by 2 and by 1, beginning with the rightmost digit. A control digit p will then be added to the sum of these products such that the new total is the next number divisible by 10. Given the account number $abcd\ efghi\ p$ with control digit p, then the congruence

$$2i + h + 2g + f + 2e + d + 2c + b + 2a + p \equiv 0 \ (\mathrm{mod}\ 10). \tag{5.265}$$

holds.

b) As in method **a)**, however, any two-digit product is first replaced by the sum of its two digits and then the total sum will be calculated.

In case **a)** all errors caused by the interchange of adjacent digits and almost all single-digit errors will be detected.
In case **b)**, however, all errors caused by the change of one digit and almost all errors caused by the interchange of two adjacent digits will be discovered. Errors due to the interchange of non-adjacent digits and the change of two digits will often not be detected.

The reason for not using the more powerful control digit method modulo 11 is of a non-mathematical nature. The non-numerical sign X (instead of the control digit 10 (see 5.4.6, **1.**, p. 383)) would require an extension of the numerical keyboard. However, renouncing those account numbers whose control digit has the value of 10 would have barred the smooth extension of the original account number in a considerable number of cases.

4. European Article Number EAN

EAN stands for *European Article Number*. It can be found on most articles as a bar code or as a string of 13 or 8 digits. The bar code can be read by means of a scanner at the counter.
In the case of 13-digit strings the first two digits identify the country of origin, e.g., $40, 41, 42$ and 43 stand for Germany. The next five digits identify the producer, the following five digits identify a particular product. The last digit is the control digit p.

This control digit will be obtained by first multiplying all 12 digits of the string alternately by 1 and 3 starting with the left-most digit, by then totalling all values, and by finally adding a p such that the next number divisible by 10 is obtained. Given the article number $abcdefghikmn\ p$ with control digit p, then the congruence

$$a + 3b + c + 3d + e + 3f + g + 3h + i + 3k + m + 3n + p \equiv 0 \ (\mathrm{mod}\ 10). \tag{5.266}$$

holds.

This control digit method always permits the detection of single digit errors in the EAN and often the detection of the interchange of two adjacent digits. The interchange of two non-adjacent digits and the

change of two digits will often not be detected.

5.4.6.2 Error correcting codes

1. Model of Data Transmission and Error Correction

At transmission of messages through noisy channels the correction of errors is often possible. The message is coded first, then after transmission the usually biased codes are corrected into the right ones, so after decoding them the original message can be recovered. That case is considered now, when the length of the words of the message is k, and the length of the coded words is n, and both of them consist of only zeros and ones. Then k is the number of *information positions* and $n-k$ is the number of *redundant positions*. Every word of the message is an element of $\mathrm{GF}(2)^k$ (see 5.3.7.4 p. 363) and every word of the code is an element of $\mathrm{GF}(2)^n$. To simplify the notation the words of the message are written in the form $a_1, a_2, \ldots, a_k$, and the words of the code in the form $c_1, c_2, \ldots, c_n$. The words of the message are not transmitted, only the words of the code are.
An often used idea of error correction is to convert the transmitted word $d_1, d_2, \ldots, d_n$ first into a valid codeword $c_1, c_2, \ldots, c_n$ which differs from it in the least number of digits (decoding MLD). It depends on the properties of coding and the transmission channels that how many errors can be detected and corrected in this way.

■ At digit repeating codes the message word 0 is represented by the codeword 0000. If after transmission the receiver gets the word 0010, then he assumes that the original codeword was 0000, and it is decoded as message word 0. But if the received word is 1010, then similar assumption can not be applied, since the message word 1 is coded as 1111, so the difference is similar. At least it can be recognized that there is some error in the received word.

2. t-Error Correcting Codes

The set of all codewords is called *code* $\mathcal{C}$. The *distance* of two codewords is the number of digits (positions) in which the two words differ from each other. The *minimal distance* $d_{\min}(\mathcal{C})$ of codes is the smallest distance which occurs between the codewords of $\mathcal{C}$.

■ For $\mathcal{C}_1 = \{0000, 1111\}$, $d_{\min}(\mathcal{C})_1 = 4$. For $\mathcal{C}_2 = \{000, 011, 101, 110\}$, $d_{\min}(\mathcal{C}_2) = 2$, since there are codewords which have distance 2. For $\mathcal{C}_3 = \{00000, 01101, 10111, 11010\}$, $d_{\min}(\mathcal{C}_3) = 3$, there are codewords in $\mathcal{C}_3$ whose distance is 3.
If the minimal distance $d_{\min}(\mathcal{C})$ of a code $\mathcal{C}$ is known, then it is easy to recognize how many transmission errors can be corrected. Codes, correcting t errors, are called *t-error correcting*. A code $\mathcal{C}$ is t-error correcting if $d_{\min}(\mathcal{C}) \geq 2t+1$.

■ (Continuation) $\mathcal{C}_1$ is 1-error correcting, $\mathcal{C}_2$ is 0-error correcting (it means, that no error can be corrected), $\mathcal{C}_3$ is 1-error correcting.
For every t-error correcting code $\mathcal{C} \subseteq \mathrm{GM}(2)^n$ holds $\sum\limits_{i=0}^{t} \binom{c}{n} \cdot |\mathcal{C}| \leq 2^n$. If equality holds, then $\mathcal{C}$ is called t-perfect.

■ The digit repeating code $\mathcal{C} = \{00\ldots0, 11\ldots1\} \subseteq \mathrm{GF}(2)^{2t+1}$ is t-perfect.

3. Linear Codes

A non-empty subset $\mathcal{C} \subseteq \mathrm{GF}(2)^n$ is called *(binary) linear code*, if $\mathcal{C}$ is a sub-vector space of $\mathrm{GF}(2)^n$. If a linear code $\mathcal{C} \subseteq \mathrm{GF}(2)^n$ has dimension k, then it is called an (n,k) *linear code*.

■ (Continuation) $\mathcal{C}_1$ is a (4,1) linear code, $\mathcal{C}_2$ is a (3,2) linear code, $\mathcal{C}_3$ is a (5,2) linear code. In the case of linear codes the minimal distance (and as a consequence the number of correctible errors) is easy to determine: The minimal distance of such a code is the smallest distance of a non-zero vector from the zero vector of the vector space. The minimal distance can be found if the minimal number of ones, except with all zeros, in the codewords is given.
For every (n,k) linear code there is a *generating matrix* $\mathbf{G}$ for which $\mathcal{C} = \{aG \mid a \in \mathrm{GF}(2)^k\}$:

$$G = \begin{pmatrix} g_{11} & \cdots & g_{1n} \\ \vdots & \vdots & \vdots \\ g_{k1} & \cdots & g_{kn} \end{pmatrix}_{k\times n} = \begin{pmatrix} g_1 \\ \vdots \\ g_k \end{pmatrix}. \tag{5.267}$$

The code is uniquely defined by the generating matrix; the codeword of the message word $a_1a_2\dots a_k$ is determined in the following way:

$$a_1a_2\dots a_k \mapsto \underbrace{a_1g_1 + a_2g_2 + \dots + a_kg_k}_{aG}. \quad (5.268)$$

In the case of an (n,k) linear code $\mathcal{C}$ a *check matrix* is needed for decoding:

$$H = \begin{pmatrix} h_{11} & \dots & h_{1n} \\ \vdots & \vdots & \vdots \\ h_{n-k,1} & \dots & h_{n-k,n} \end{pmatrix}_{(n-k)\times n}. \quad (5.269)$$

The (binary) linear code $\mathcal{C}$ is 1-error correcting, if the columns of H are pairwise different and non-zero vectors. If the result of the transmission is the word $d = d_1d_2\dots d_n$, then Hd^T is calculated. If the result is the zero vector, then d is a codeword. Otherwise if Hd^T is the i-th column of the check matrix H, then the corresponding codeword is $d + e_i$, where $e_i = (0, 0, \dots, 0, 1, 0, \dots, 0)$ and the 1 is on the i-th position.

4. Cyclic Codes

Cyclic codes are the most investigated linear codes. They provide efficient coding and decoding.
A (binary) (n,k) linear code is called *cyclic* if for every codeword $c_1c_2\dots c_n$ the codeword obtained by a cyclic right shift of the components is also a codeword, i.e. $c_0c_1\dots c_{n-1}\in\mathcal{C} \Rightarrow c_{n-1}c_0c_1\dots c_{n-2}\in\mathcal{C}$

■ $\mathcal{C} = \{000, 110, 101, 011\}$ is a cyclic (3,2) linear code.

To have an efficient work with cyclic codes, the codewords are represented by polynomials of degree $\le n-1$ with coefficients from GF(2): $\mathcal{C} = \{000, 110, 101, 011\}$ is a cyclic $(3,2)$-linear code.
A (binary) (n,k) linear code $\mathcal{C}$ is cyclic if and only if for every $c(x)$

$$c(x)\in\mathcal{C} \Rightarrow c(x)\cdot x (\mathrm{mod}\ x^n\ \text{-}1)\in\mathcal{C} \quad (5.270)$$

A cyclic (n,k) linear code can be described by a generating polynomial and a control polynomial as follows: The *generating polynomial* $g(x)$ of degree $n-k$ ($k\in\{1,2,\dots,n-1\}$) is a divisor of x^n-1. The polynomial $h(x)$ of degree k for which $g(x)h(x) = x^n - 1$ is called the *control polynomial*. Coding of $a_1a_2\dots a_k$ in polynomial representation $a(x)$ is given by

$$a(x)\mapsto a(x)\cdot g(x). \quad (5.271)$$

Polynomial $d(x)$ is an element of the code, if the generator polynomial $g(x)$ is a divisor of $d(x)$, or the control polynomial $h(x)$ satisfies the relation $d(x)h(x)\equiv 0 \bmod x^n - 1$.

An important class of cyclic codes are the BCH-codes. Here a lower bound δ of the minimal distance and with it a lower bound for the number of errors can be required for which code should be corrected. Here δ is called the *design distance* of the code.
A (binary) (n,k) linear code $\mathcal{C}$ is a BCH-code with design distance δ if for the generating polynomial $g(x)$:

$$g(x) = \mathrm{lcm}(m_{\alpha^b}(x), m_{\alpha^{b+1}}(x), \dots, m_{\alpha^{b+\delta-2}}(x)), \quad (5.272)$$

where α is a primitive n-th unit root and b is an integer. The polynomials $m_{\alpha^j}(x)$ are minimal polynomials of α^j.
For a BCH-code $\mathcal{C}$ with design distance δ the relation $d_{\min}(\mathcal{C})\ge\delta$ must hold.

5.5 Cryptology

5.5.1 Problem of Cryptology

Cryptology is the science of hiding information by the transformation of data.

The idea of protecting data from unauthorized access is rather old. During the 1970s together with the introduction of *cryptosystems on the basis of public keys*, cryptology became an independent branch of science. Today, the subject of cryptological research is how to protect data from unauthorized access and against tampering.

Beside the classical military applications, the needs of the information society gain more and more in importance. Examples are the guarantee of secure message transfer via email, electronic funds transfer (home-banking), the PIN of EC-cards, etc.

Today, the fields of *cryptography* and *cryptanalysis* are subsumed under the notion of cryptology. Cryptography is concerned with the development of cryptosystems whose cryptographic strengths can be assessed by applying the methods of cryptanalysis for breaking cryptosystems.

5.5.2 Cryptosystems

An abstract cryptosystem consists of the following sets: a set M of messages, a set C of ciphertexts, sets K and K' of keys, and sets $\mathbb{E}$ and $\mathbb{D}$ of functions. A message $m \in M$ will be encrypted into a ciphertext $c \in C$ by applying a function $E \in \mathbb{E}$ together with a key $k \in K$, and will be transmitted via a communication channel. The recipient can reproduce the original message m from c if he knows an appropriate function $D \in \mathbb{D}$ and the corresponding key $k' \in K'$. There are two types of cryptosystems:

1. Symmetric Cryptosystems: The conventional symmetric cryptosystem uses the same key k for encryption of the message and for decryption of the ciphertext. The user has complete freedom in setting up his conventional cryptosystem. Encryption and decryption should, however, not become too complex. In any case, a trustworthy transmission between the two communication partners is mandatory.

2. Asymmetric Cryptosystems: The asymmetric cryptosystem (see 5.5.7.1, p. 391) uses two keys, one private key (to be kept secret) and a public key. The public key can be transmitted along the same path as the ciphertext. The security of the communication is warranted by the use of so-called *one-way functions* (see 5.5.7.2, p. 391), which makes it practically impossible for the unauthorized listener to deduce the plaintext from the ciphertext.

5.5.3 Mathematical Foundation

An alphabet $A = \{a_0, a_1, \ldots, a_{n-1}\}$ is a finite non-empty totally ordered set, whose elements a_i are called letters. $|A|$ is the length of the alphabet. A sequence of letters $w = a'_1 a'_2 \ldots a'_n$ of length $n \in \mathbb{N}$ and $a_i \in A$ is called a word of length n over the alphabet A. A^n denotes the set of all words of length n over A. Let $n, m \in \mathbb{N}$, let A, B be alphabets, and let S be a finite set.

A cryptofunction is a mapping t: $A^n \times S \to B^m$ such that the mappings t_s: $A^n \to B^m : w \to t(w, s)$ are injective for all $s \in S$. The functions t_s and t_s^{-1} are called the encryption and decryption function, respectively. w is called plaintext, $t_s(w)$ is the ciphertext.

Given a cryptofunction t, then the one-parameter family $\{t_s\}_{s\in S}$ is a cryptosystem T_S. The term *cryptosystem* will be applied if in addition to the mapping t, the structure and the size of the set of keys is significant. The set S of all the keys belonging to a cryptosystem is called the key space. Then

$$T_S = \{t_s\colon A^n \to A^n | s \in S\} \tag{5.273}$$

is called a cryptosystem on A^n.

If T_S is a cryptosystem over A^n and $n = 1$, then t_s is called a stream cipher; otherwise t_s is called a block cipher.

Cryptofunctions of a cryptosystem over A^n are suited for the encryption of plaintext of any length. The plaintext will be split into blocks of length n prior to applying the function to each individual block. The last block may need padding with filler characters to obtain a block of length n. The filler characters must not distort the plaintext.

There is a distinction between *context-free encryption*, where the ciphertext block is only a function of the corresponding plaintext block and the key, and *context sensitive encryption*, where the ciphertext block depends on other blocks of the message. Ideally, each ciphertext digit of a block depends on all digits of the corresponding plaintext block and all digits of the key. Small changes to the plaintext or

to the key cause extended changes to the ciphertext (avalanche effect).

5.5.4 Security of Cryptosystems

Cryptanalysis is concerned with the development of methods for deducing from the ciphertext as much information about the plaintext as possible without knowing the key. According to A. Kerkhoff the security of a cryptosystem rests solely in the difficulty of detecting the key or, more precisely, the decryption function. The security must not be based on the assumption that the encryption algorithm is kept secret. There are different approaches to assess the security of a cryptosystem:

1. Absolutely Secure Cryptosystems: There is only one absolutely secure cryptosystem based on substitution ciphers, which is the *one-time pad*. This was proved by Shannon as part of his information theory.

2. Analytically Secure Cryptosystems: No method exists to break a cryptosystem systematically. The proof of the non-existence of such a method follows from the proof of the non-computability of a decryption function.

3. Secure Cryptosystems according to Criteria of Complexity Theory: There is no algorithm which can break a cryptosystem in polynomial time (with regard to the length of the text).

4. Practically Secure Cryptosystems: No method is known which can break the cryptosystem with available resources and with justified costs.

Cryptanalysis often applies statistical methods such as determining the frequency of letters and words. Other methods are an exhaustive search, the trial-and-error method and a structural analysis of the cryptosystem (solving of equation systems).

In order to attack a cryptosystem one can benefit from frequent flaws in encryption such as using stereotype phrases, repeated transmissions of slightly modified text, an improper and predictable selection of keys, and the use of filler characters.

5.5.4.1 Methods of Conventional Cryptography

In addition to the application of a cryptofunction it is possible to encrypt a plaintext by means of *cryptological codes*. A *code* is a bijective mapping of some subset A' of the set of all words over an alphabet A onto the subset B' of the set of all words over the alphabet B. The set of all source-target pairs of such a mapping is called a code book.

■

today evening	0815
tomorrow evening	1113

The advantage of replacing long plaintexts by short ciphertexts is contrasted with the disadvantage that the same plaintext will always be replaced by the same ciphertext. Another disadvantage of code books is the need for a complete and costly replacement of all books should the code be compromised even partially.

In the following only encryption by means of cryptofunctions will be considered. Cryptofunctions have the additional advantage that they do not require any arrangement about the contents of the messages prior to their exchange.

Transposition and *substitution* constitute conventional cryptoalgorithms. In cryptography, a transposition is a special permutation defined over geometric patterns. The substitutions will now be discussed in detail. There is a distinction between monoalphabetic and polyalphabetic substitutions according to how many alphabets are used for presenting the ciphertext. Generally, a substitution is termed polyalphabetic even if only one alphabet is used, but the encryption of the individual plaintext letter depends on its position within the plaintext.

A further, useful classification is the distinction between monographic and polygraphic substitutions. In the first case, single letters will be substituted, in the latter case, strings of letters of a fixed length > 1.

5.5.4.2 Linear Substitution Ciphers

Let $A = \{a_0, a_1, \dots, a_{n-1}\}$ be an alphabet and $k, s \in \{0, 1, \dots, n-1\}$ with $\gcd(k, n) = 1$. The permutation t_s^k, which maps each letter a_i to $t_s^k(a_i) = a_{ki+s}$, is called a linear substitution cipher. There exist $n\,\varphi(n)$ linear substitution ciphers on A.

Shift ciphers are linear substituting ciphers with $k = 1$. The shift cipher with $s = 3$ was already used by Julius Caesar (100 to 44 BC) and, therefore, it is called the Caesar cipher.

5.5.4.3 Vigenère Cipher

An encryption called the Vigenère cipher is based on the periodic application of a key word whose letters are pairwise distinct. The encryption of a plaintext letter is determined by the key letter that has the same position in the key as the plaintext letter in the plaintext. This requires a key that is as long as the plaintext. Shorter keys are repeated to match the length of the plaintext.

A version of the Vigenère cipher attributed to L. Carroll utilizes the so-called Vigenère tableau (see picture) for encryption and decryption. Each row represents the cipher for the key letter to its very left. The alphabet for the plaintext runs across the top. The encryption step is as follows: Given a key letter D and a plaintext letter C, then the ciphertext letter is found at the intersection of the row labeled D and the column labeled C; the ciphertext is F. Decryption is the inverse of this process.

	A	B	C	D	E	F	...
A	A	B	C	D	E	F	...
B	B	C	D	E	F	G	...
C	C	D	E	F	G	H	...
D	D	E	F	G	H	I	...
E	E	F	G	H	I	J	...
F	F	G	H	I	J	K	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱

■ Let the key be " HUT ".

Plaintext:	O	N	C	E	U	P	O	N	A	T	I	M	E
Key:	H	U	T	H	U	T	H	U	T	H	U	T	H
Ciphertext:	V	H	V	L	O	I	V	H	T	A	C	F	L

Formally, the Vigenère cipher can be written in the following way: let a_i be the plaintext letter and a_j be the corresponding key letter, then $k = i + j$ determines the ciphertext letter a_k. In the above example, the first plaintext letter is $O = a_{14}$. The 15-th position of the key is taken by the letter $H = a_7$. Hence, $k = i + j = 14 + 7 = 21$ yields the ciphertext letter $a_{21} = V$.

5.5.4.4 Matrix Substitution

Let $A = \{a_0, a_1, \dots, a_{n-1}\}$ be an alphabet and $S = (s_{ij})$, $s_{ij} \in \{0, 1, \dots, m-1\}$, be a non-singular matrix of type (m, m) with $\gcd(\det S, n) = 1$. The mapping which maps the block of plaintext $a_{t(1)}$, $a_{t(2)}, \dots, a_{t(m)}$ to the ciphertext determined by the vector (all arithmetic modulo n, vectors transposed as required)

$$\left(S \cdot \begin{pmatrix} a_{t(1)} \\ a_{t(2)} \\ \vdots \\ a_{t(m)} \end{pmatrix} \right)^T \tag{5.274}$$

is called the Hill cipher. This represents a monoalphabetic matrix substitution.

■ $S = \begin{pmatrix} 14 & 8 & 3 \\ 8 & 5 & 2 \\ 3 & 2 & 1 \end{pmatrix}$. Let the letters of the alphabet be enumerated $a_0 =$ A, $a_1 =$B, $\dots, a_{25} =$ Z. For $m = 3$ and the plaintext AUTUMN, the strings AUT and UMN correspond to the vectors $(0, 20, 19)$ and $(20, 12, 13)$.

Then $S \cdot (0, 20, 19)^\top = (217, 138, 59)^\top \equiv (9, 8, 7)^\top (\text{mod} 26)$ and $S \cdot (20, 12, 13)^\top = (415, 246, 97)^\top \equiv (25, 12, 19)^\top (\text{mod} 26)$. Thus, the plaintext AUTUMN is mapped to the ciphertext JIHZMT.

5.5.5 Methods of Classical Cryptanalysis

The purpose of cryptanalytical investigations is to deduce from the ciphertext an optimum of information about the corresponding plaintext without knowing the key. These analyses are of interest not

only to an unauthorized "eavesdropper" but also help assess the security of cryptosystems from the user's point of view.

5.5.5.1 Statistical Analysis

Each natural language shows a typical frequency distribution of the individual letters, two-letter combinations, words, etc. For example, in English the letter e is used most frequently:

Letter	Relative frequency
E,	12.7 %
T, A, O, I, N, S, H, R	56.9 %
D, L	8.3 %
C, U, M, W, F, G, Y, P, B	19.9 %
V, K, J, X, Q, Z	2.2 %

Given sufficiently long ciphertexts it is possible to break a monoalphabetic, monographic substitution on the basis of the frequency distribution of letters.

5.5.5.2 Kasiski-Friedman Test

Combining the methods of Kasiski and Friedman it is possible to break the Vignère cipher. The attack benefits from the fact that the encryption algorithm applies the key periodically. If the same string of plaintext letters is encrypted with the same portion of the key then the same string of ciphertext letters will be produced. A length > 2 of the distance of such identical strings in the ciphertext must be a multiple of the key length. In the case of several reoccurring strings of ciphertext the key length is a divisor of the greatest common divisor of all distances. This reasoning is called the Kasiski test. One should, however, be aware of erroneous conclusions due to the possibility that matches may occur accidentally.

The Kasiski test permits the determination of the key length at most as a multiple of the true key length. The Friedman test yields the magnitude of the key length. Let n be the length of the ciphertext of some English plaintext encrypted by means of the Vignère method. Then the key length l is determined by

$$l = \frac{0.027n}{(n-1)\mathrm{IC} - 0.038\mathrm{n} + 0.065}. \tag{5.275a}$$

Here IC denotes the coincidence index of the ciphertext. This index can be deduced from the number n_i of occurrences of the letter a_i $(i \in \{0, 1, \ldots, 25\})$ in the ciphertext:

$$\mathrm{IC} = \frac{\sum\limits_{\mathrm{i}=1}^{26} \mathrm{n_i}(\mathrm{n_i} - 1)}{\mathrm{n}(\mathrm{n} - 1)}. \tag{5.275b}$$

In order to determine the key, the ciphertext of length n is split into l columns. Since the Vignère cipher produces the contents of each column by means of a shift cipher, it suffices to determine the equivalence of E on a column base. Should V be the most frequent letter within a column, then the Vignère tableau points to the letter R

$$\begin{array}{cc} & \mathrm{E} \\ & \vdots \\ \mathrm{R}\ldots & \mathrm{V} \end{array} \tag{5.275c}$$

of the key. The methods described so far will not be successful if the Vignère cipher employs very long keys (e.g., as long as the plaintext). It is, however, possible to deduce whether the applied cipher is monoalphabetic, polyalphabetic with short period or polyalphabetic with long period.

5.5.6 One-Time Pad

The one-time pad is the only substitution cipher that is considered theoretically secure. The encryption adheres to the principle of the Vignère cipher, where the key is a random string of letters as long as the

plaintext.

Usually, one-time pads are applied as binary Vignère ciphers: Plaintext and ciphertext are represented as binary numbers with addition modulo 2. In this particular case the cipher is involutory, which means that the twofold application of the cipher restores the original plaintext. A concrete implementation of the binary Vignère cipher is based on shift register circuits. These circuits combine switches and storage elements, whose states are 0 or 1, according to special rules.

5.5.7 Public Key Methods

Although the methods of conventional encryption can have efficient implementations with today's computers, and although only a single key is needed for bidirectional communication, there are a number of drawbacks:

1. The security of encryption solely depends on keeping the next key secret.

2. Prior to any communication, the key must be exchanged via a sufficiently secured channel; spontaneous communication is ruled out.

3. Furthermore, no means exist to prove to a third party that a specific message was sent by an identified sender.

5.5.7.1 Diffie-Hellman Key Exchange

The concept of encryption with public keys was developed by Diffie and Hellman in 1976. Each participant owns two keys: a public key that is published in a generally accessible register, and a private key that is solely known to the participant and kept absolutely secret. Methods with these properties are called asymmetric ciphers (see 5.5.2, p. 387).

The public key KP_i of the i-th participant controls the encryption step E_i, his private key KS_i the decryption step D_i. The following conditions must be fulfilled:

1. $D_i \circ E_i$ constitutes the identity.

2. Efficient implementations for E_i and D_i are known.

3. The private key KS_i cannot be deduced from the public key KP_i with the means available in the foreseeable future. If in addition

4. also $E_i \circ D_i$ yields the identity,

then the encryption algorithm qualifies as an electronic signature method with public keys. The electronic signature method permits the sender to attach a tamperproof signature to a message.

If A wants to send an encrypted message m to B, then A retrieves B's public key KP_B from the register, applies the encryption algorithm E_B, and calculates $E_B(m) = c$. A sends the ciphertext c via the public network to B who will regain the plaintext of the message by decrypting c using his private key KS_B in the decryption function D_B: $D_B(c) = D_B(E_B(m)) = m$. In order to prevent tampering of messages, A can electronically sign his message m to B by complying with an electronic signature method with the public key in the following way: A encrypts the message m with his private key: $D_A(m) = d$. A attaches to d his signature "A" and encrypts the total using the public key of B: $E_B(D_A(m), \text{"}A\text{"}) = E_B(d, \text{"}A\text{"}) = e$. The text thus signed and encrypted is sent from A to B.

The participant B decrypts the message with his private key and obtains $D_B(e) = D_B(E_B(d, \text{"}A\text{"})) = (d, \text{"}A\text{"})$. Based on this text B can identify A as the sender and can now decrypt d using the public key of $A : E_A(d) = E_A(D_A(m)) = m$.

5.5.7.2 One-Way Function

The encryption algorithms of a method with public key must constitute a one-way function with a "trap door". A trap door in this context is some special, additional information that must be kept secret. An injective function $f\colon X \longrightarrow Y$ is called a one-way function with a trap door, if the following conditions hold:

1. There is an efficient method to compute both f and f^{-1}.

2. The calculation of f^{-1} cannot be deduced from f without the knowledge of the secret additional information.
The efficient method to get f^{-1} from f cannot be made without the secret additional information.

5.5.7.3 RSA Codes and RSA Method

1. RSA Codes

Rivest, Shamir and Adleman (see [5.16]) developed an encryption scheme for secret messages on the basis of the Euler-Fermat theorem (see 5.4.4, **2.**, p. 381). The scheme is called the *RSA algorithm* after the initials of their last names. Part of the key required for decryption can be made public without endangering the confidentiality of the message; for this reason, the term *public key code* is used in this context as well.
In order to apply the RSA algorithm the recipient B chooses two very large prime numbers p and q, calculates $m = pq$ and selects a number r relatively prime to $\varphi(m) = (p-1)(q-1)$ and $1 < r < \varphi(m)$. B publishes the numbers m and r because they are needed for decryption.
For transmitting a secret message from sender A to recipient B the text of the message must be converted first to a string of digits that will be split into N blocks of the same length of less than 100 decimal positions. Now A calculates the remainder R of N^r divided by m.

$$N^r \equiv R(m). \tag{5.276a}$$

Sender A calculates the number R for each of the blocks N that were derived from the original text and sends the number to B. The recipient can decipher the message R if he has a solution of the linear congruence $rs \equiv 1\,(\varphi(m))$. The number N is the remainder of R^s divided by m:

$$R^s \equiv (N^r)^s \equiv N^{1+k\varphi(m)} \equiv N \cdot (N^{\varphi(m)})^k \equiv N(m). \tag{5.276b}$$

Here, the Euler-Fermat theorem (see 5.4.4, **2.**, p. 381) with $N^{\varphi(m)} \equiv 1(m)$ has been applied. Eventually, B converts the sequence of numbers into text.

■ A recipient B who expects a secret message from sender A chooses the prime numbers $p = 29$ and $q = 37$ (actually too small for practical purposes), calculates $m = 29 \cdot 37 = 1073$ (and $\varphi(1073) = \varphi(29)\cdot\varphi(37) = 1008$)), and chooses $r = 5$ (it satisfies the requirement of $\gcd(1008, 5) = 1$). B passes the values $m = 1073$ and $r = 5$ to A.
A intends to send the secret message $N = 8$ to B. A encrypts N into $R = 578$ by calculating $N^r = 8^5 \equiv 578\,(1073)$, and just sends the value $R = 578$ to B. B solves the congruence $5 \cdot s \equiv 1\,(1008)$, arrives at the solution $s = 605$, and thus determines $R^s = 578^{605} \equiv 8 = N\,(1073)$.

Remark: The security of the RSA code correlates with the time needed by an unauthorized listener to factorize m. Assuming the speed of today's computers, a user of the RSA algorithm should choose the two prime numbers p and q with at least a length of 100 decimal positions in order to impose a decryption effort of approximately 74 years on the unauthorized listener. The effort for the authorized user, however, to determine an r relatively prime to $\varphi(pq) = (p-1)(q-1)$ is comparatively small.

2. RSA Method

The RSA method is the most popular asymmetric encryption method.

1. Assumptions Let p and q be two large prime numbers with $pq \approx 10^{2048}$ and $n = pq$. The number of decimal positions of p and q should differ by a small number; yet, the difference between p and q should not be too large. Furthermore, the numbers $p-1$ and $q-1$ should contain rather big prime factors, while the greatest common divisor of $p-1$ and $q-1$ should be rather small. Let $e > 1$ be relatively prime to $(p-1)(q-1)$ and let d satisfy $d \cdot e \equiv 1\,(\mathrm{mod}(p-1)(q-1))$. Now n and e represent the public key and d the private key.

2. Encryption Algorithm

$$E\colon \{0, 1, \ldots, n-1\} \to \{0, 1, \ldots, n-1\} \quad E(x) := x^e \bmod n. \tag{5.277a}$$

3. Decyphering Operations

$$D\colon \{0, 1, \ldots, n-1\} \to \{0, 1, \ldots, n-1\} \quad D(x) := x^d \bmod n. \tag{5.277b}$$

Thus $D(E(m)) = E(D(m) = m$ for message m.

The function in this encryption method with $n > 10^{200}$ constitutes a candidate for a one-way function with trap door (see 5.5.7.2, p. 391). The required additional information is the knowledge of how to factor n. Without this knowledge it is infeasible to solve the congruence $d \cdot e \equiv 1$ (modulo $(p-1)(q-1)$).

The RSA method is considered practically secure as long as the above conditions are met. A disadvantage in comparison with other methods is the relatively large key size and the fact that RSA is 1000 times slower than DES.

5.5.8 DES Algorithm (Data Encryption Standard)

The DES method was adopted in 1976 by the National Bureau of Standards (now NIST) as the official US encryption standard. The algorithm belongs to the class of symmetric encryption methods (see 5.5.2, p. 387) and still plays a predominant role among cryptographic methods. The method is, however, no longer suited for the encryption of top secret information because today's technical means permit an attack by an exhaustive test trying all keys.

The DES algorithm combines permutations and non-linear substitutions. The algorithm requires a 56-bit key. Actually, a 64-bit key is used, however, only 56 bits can freely be chosen; the remaining eight bits serve as parity bits, one for each of the seven-bit blocks to yield odd parity.

The plaintext is split into blocks of 64 bits each. DES transforms each 64-bit plaintext block into a ciphertext block of 64 bits. First, the plaintext block will be subject to an initial permutation and is then encrypted in 16 rounds, each operating with a different subkey $K_1, K_2, \ldots, K_{16}$. The encryption completes with a final permutation that is the inverse of the initial permutation.

Decryption uses the same algorithm with the difference that the subkeys are employed in reverse order $K_{16}, K_{15}, \ldots, K_1$.

The strength of the cipher rests on the nature of the mappings that are part of each round. It can be shown that each bit of the ciphertext block depends on each bit of the corresponding plaintext and on each bit of the key.

Although the DES algorithm has been disclosed in full detail, no attack has been published so far that can break the algorithm without an exhaustive test of all 256 keys.

5.5.9 IDEA Algorithm (International Data Encryption Algorithm)

The IDEA algorithm was developed by LAI and MASSAY and patented 1991. It is a symmetric encryption method similar to the DES algorithm and constitutes a potential successor to DES. IDEA became known as part of the reputed software package PGP (Pretty Good Privacy) for the encryption of emails. In contrast to DES not only was the algorithm published but even its basic design criteria. The objective was the use of particularly simple operations (addition modulo 2, addition modulo 2^{16}, multiplication modulo 2^{16+1}).

IDEA works with keys of 128 bits length. IDEA encrypts plaintext blocks of 64 bits each. The algorithm splits a block into four subblocks of 16 bits each. From the 128-bit key 52 subkeys are derived, each 16 bits long. Each of the eight encryption rounds employs six subkeys; the remaining four subkeys are used in the final transformation which constructs the resulting 64-bit ciphertext. Decryption uses the same algorithm with the subkeys in reverse order.

IDEA is twice as fast as DES, its implementation in hardware, however, is more difficult. No successful attack against IDEA is known. Exhaustive attacks trying all 2^{56} keys are infeasible considering the length of the keys.

5.6 Universal Algebra

A *universal algebra* consists of a set, the *underlying set*, and operations on this set. Simple examples are semigroups, groups, rings, and fields discussed in sections 5.3.2, p. 336; 5.3.3, p. 336 and 5.3.7, p. 361. Universal algebras (mostly many-sorted, i.e., with several underlying sets) are handled especially in theoretical informatics. There they form the basis of algebraic specifications of abstract data types and systems and of term-rewriting systems.

5.6.1 Definition

Let Ω be a set of operation symbols divided into pairwise disjoint subsets Ω_n, $n \in \mathbb{N}$. Ω_0 contains the constants, Ω_n, $n > 0$, contain the n-ary operation symbols. The family $(\Omega_n)_{n\in\mathbb{N}}$ is called the *type* or *signature*. If A is a set, and if to every n-ary operation symbol $\omega \in \Omega_n$ an n-ary operation ω^A in A is assigned, then $A = (A, \{\omega^A | \omega \in \Omega\})$ is called an Ω *algebra* or algebra of type (or of signature) Ω.
If Ω is finite, $\Omega = \{\omega_1, \ldots, \omega_k\}$, then one also writes $A = (A, \omega_1^A, \ldots, \omega_k^A)$ for A.
If a ring (see 5.3.7, p. 361) is considered as an Ω algebra, then Ω is partitioned $\Omega_0 = \{\omega_1\}$, $\Omega_1 = \{\omega_2\}$, $\Omega_2 = \{\omega_3, \omega_4\}$, where to the operation symbols ω_1, ω_2, ω_3, ω_4 the constant 0, taking the inverse with respect to addition, addition and multiplication are assigned.

Let A and B be Ω algebras. B is called an Ω *subalgebra* of A, if $B \subseteq A$ holds and the operations ω^B are the restrictions of the operations ω^A $(\omega \in \Omega)$ to the subset B.

5.6.2 Congruence Relations, Factor Algebras

In constructing factor structures for universal algebras, the notion of congruence relation is needed. A congruence relation is an equivalence relation compatible with the structure: Let $A = (A, \{\omega^A | \omega \in \Omega\})$ be an Ω algebra and R be an equivalence relation in A. R is called a *congruence relation* in A, if for all $\omega \in \Omega_n$ $(n \in \mathbb{N})$ and all $a_i, b_i \in A$ with $a_i R b_i$ $(i = 1, \ldots, n)$:

$$\omega^A(a_1, \ldots, a_n) \; R \; \omega^A(b_1, \ldots, b_n). \tag{5.278}$$

The set of equivalence classes (factor set) with respect to a congruence relation also form an Ω algebra with respect to representative-wise calculations: Let $A = (A, \{\omega^A | \omega \in \Omega\})$ be an Ω algebra and R be a congruence relation in A. The factor set A/R (see 5.2.4, **2.**, p. 334) is an Ω algebra A/R with the following operations $\omega^{A/R}$ $(\omega \in \Omega_n, n \in \mathbb{N})$ with

$$\omega^{A/R}([a_1]_R, \ldots, [a_n]_R) = [\omega^A(a_1, \ldots, a_n)]_R \tag{5.279}$$

and it is called the *factor algebra* of A with respect to R.

The congruence relations of groups and rings can be defined by special substructures – normal subgroups (see 5.3.3.2, **2.** p. 338) and ideals (see 5.3.7.2, p. 362), respectively. In general, e.g., in semigroups, such a characterization of congruence relations is not possible.

5.6.3 Homomorphism

Just as with classical algebraic structures, the homomorphism theorem gives a connection between the homomorphisms and congruence relations.
Let A and B be Ω algebras. A mapping h: $A \to B$ is called a *homomorphism*, if for every $\omega \in \Omega_n$ and all $a_1, \ldots, a_n \in A$:

$$h(\omega^A(a_1, \ldots, a_n)) = \omega^B(h(a_1), \ldots, h(a_n)). \tag{5.280}$$

If, in addition, h is bijective, then h is called an *isomorphism*; the algebras A and B are called *isomorphic*. The homomorphic image $h(A)$ of an Ω algebra A is an Ω subalgebra of B. Under a homomorphism h, the decomposition of A into subsets of elements with the same image corresponds to a congruence relation which is called the *kernel* of h:

$$\ker\ h = \{(a, b) \in A \times A | h(a) = h(b)\}. \tag{5.281}$$

5.6.4 Homomorphism Theorem

Let A and B be Ω algebras and h: $A \to B$ a homomorphism. h defines a congruence relation ker h in A. The factor algebra $A/\ker\ h$ is isomorphic to the homomorphic image $h(A)$.
Conversely, every congruence relation R defines a homomorphic mapping nat_R: $A \to A/R$ with $nat_R(a) = [a]_R$. **Fig. 5.19** illustrates the homomorphism theorem.

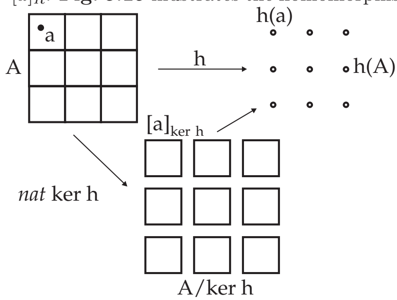


Figure 5.19

5.6.5 Varieties

A *variety* V is a class of Ω algebras, which is closed under forming direct products, subalgebras, and homomorphic images, i.e., these formations do not lead out of V. Here the direct products are defined in the following way:

Considering the operations corresponding to Ω componentwise on the Cartesian product of the underlying sets of Ω algebras, an Ω algebra, the *direct product* of these algebras is obtained. The theorem of Birkhoff (see 5.6.6, p. 395) characterizes the varieties as those classes of Ω algebras, which can be *equationally defined*.

5.6.6 Term Algebras, Free Algebras

Let $(\Omega_n)_{n\in\mathbb{N}}$ be a type (signature) and X a countable set of variables. The set $T_\Omega(X)$ of Ω terms over X is defined inductively in the following way:

1. $X \cup \Omega_0 \subseteq T_\Omega(X)$.

2. If $t_1, \ldots, t_n \in T_\Omega(X)$ and $\omega \in \Omega_n$ hold, then also $\omega t_1 \ldots t_n \in T_\Omega(X)$ holds.

The set $T_\Omega(X)$ defined in this way is an underlying set of an Ω algebra, the *term algebra* $T_\Omega(X)$ of type Ω over X, with the following operations: If $t_1, \ldots, t_n \in T_\Omega(X)$ and $\omega \in \Omega_n$ hold, then $\omega^{T_\Omega(X)}$ is defined by

$$\omega^{T_\Omega(X)}(t_1, \ldots, t_n) = \omega t_1 \ldots t_n. \tag{5.282}$$

Term algebras are the "most general" algebras in the class of all Ω algebras, i.e., no "identities" are valid in term algebras. These algebras are called *free algebras*.
An *identity* is a pair $(s(x_1, \ldots, x_n), t(x_1, \ldots, x_n))$ of Ω terms in the variables $x_1, \ldots, x_n$. An Ω algebra A *satisfies* such an equation, if for every $a_1, \ldots, a_n \in A$ holds:

$$s^A(a_1, \ldots, a_n) = t^A(a_1, \ldots, a_n). \tag{5.283}$$

A class of Ω algebras defined by identities is a class of Ω algebras satisfying a given set of identities.

Theorem of Birkhoff: The classes defined by identities are exactly the varieties.

■ Varieties are for example the classes of all semigroups, groups, Abelian groups, and rings. But, e.g., the direct product of cyclic groups is not a cyclic group, and the direct product of fields is not a field. Therefore cyclic groups or fields do not form a variety, and cannot be defined by equations.

5.7 Boolean Algebras and Switch Algebra

Calculating rules, similar to the rules established in 5.2.2, **3.**, p. 329 for set algebra and propositional calculus (5.1.1, **6.**, p. 324), can be found for other objects in mathematics too. The investigation of these rules yields the notion of Boolean algebra.

5.7.1 Definition

A set B, together with two binary operations $\sqcap$ ("conjunction") and $\sqcup$ ("disjunction"), and a unary operation ("negation"), and two distinguished (neutral) elements 0 and 1 from B, is called a *Boolean*

algebra $B = (B, \sqcap, \sqcup, \bar{\ }, 0, 1)$ if the following properties are valid:

(1) Associative Laws:

$$(a \sqcap b) \sqcap c = a \sqcap (b \sqcap c), \quad (5.284) \qquad (a \sqcup b) \sqcup c = a \sqcup (b \sqcup c). \quad (5.285)$$

(2) Commutative Laws:

$$a \sqcap b = b \sqcap a, \quad (5.286) \qquad a \sqcup b = b \sqcup a. \quad (5.287)$$

(3) Absorption Laws:

$$a \sqcap (a \sqcup b) = a, \quad (5.288) \qquad a \sqcup (a \sqcap b) = a. \quad (5.289)$$

(4) Distributive Laws:

$$(a \sqcup b) \sqcap c = (a \sqcap c) \sqcup (b \sqcap c), \quad (5.290) \qquad (a \sqcap b) \sqcup c = (a \sqcup c) \sqcap (b \sqcup c). \quad (5.291)$$

(5) Neutral Elements:

$$a \sqcap 1 = a, \quad (5.292) \qquad a \sqcup 0 = a, \quad (5.293)$$

$$a \sqcap 0 = 0, \quad (5.294) \qquad a \sqcup 1 = 1, \quad (5.295)$$

(6) Complement:

$$a \sqcap \overline{a} = 0, \quad (5.296) \qquad a \sqcup \overline{a} = 1. \quad (5.297)$$

A structure with the associative laws, commutative laws, and absorption laws is called a *lattice*. If the distributive laws also hold, then the lattice is called a *distributive lattice*. So a Boolean algebra is a special distributive lattice.

Remark: The notation used for Boolean algebras is not necessarily identical to the notation for the operations in propositional calculus.

5.7.2 Duality Principle

1. Dualizing

In the "axioms" of a Boolean algebra is included the following duality: Replacing $\sqcap$ by $\sqcup$, $\sqcup$ by $\sqcap$, 0 by 1, and 1 by 0 in an axiom gives always the other axiom in the same row. The axioms in a row are *dual* to each other, and the substitution process is called *dualization*. The *dual statement* follows from a statement of the Boolean algebra by dualization.

2. Duality Principle for Boolean Algebras

The dual statement of a true statement for a Boolean algebra is also a true statement for the Boolean algebra, i.e., with every proved proposition, the dual proposition is also proved.

3. Properties

One gets, e.g., the following properties for Boolean algebras from the axioms.

(E1) The Operations $\sqcap$ and $\sqcup$ are Idempotent:

$$a \sqcap a = a, \quad (5.298) \qquad a \sqcup a = a. \quad (5.299)$$

(E2) De Morgan Rules:

$$\overline{a \sqcap b} = \overline{a} \sqcup \overline{b}, \quad (5.300) \qquad \overline{a \sqcup b} = \overline{a} \sqcap \overline{b}, \quad (5.301)$$

(E3) A further Property:

$$\overline{\overline{a}} = a. \quad (5.302)$$

It is enough to prove only one of the two properties in any line above, because the other one is the dual property. The last property is self-dual.

5.7.3 Finite Boolean Algebras

All finite Boolean algebras can be described easily up to "isomorphism". Let B_1, B_2 be two Boolean algebras and $f\colon B_1 \to B_2$ a bijective mapping. f is called an *isomorphism* if

$$f(a \sqcap b) = f(a) \sqcap f(b), \quad f(a \sqcup b) = f(a) \sqcup f(b) \quad \text{and} \quad f(\overline{a}) = \overline{f(a)} \tag{5.303}$$

hold. Every finite Boolean algebra is isomorphic to the Boolean algebra of the power set of a finite set. In particular every finite Boolean algebra has 2^n elements, and every two finite Boolean algebras with the same number of elements are isomorphic.

Hereafter B denotes the Boolean algebra with two elements $\{0, 1\}$ and with the operations

$\sqcap$	0	1
0	0	0
1	0	1

$\sqcup$	0	1
0	0	1
1	1	1

	$-$
0	1
1	0

Defining the operations $\sqcap$, $\sqcup$, and $\bar{\ }$ componentwise on the n-times Cartesian product $B^n = \{0, 1\} \times \cdots \times \{0, 1\}$, then B^n will be a Boolean algebra with $0 = (0, \ldots, 0)$ and $1 = (1, \ldots, 1)$. B^n is called the *n times direct product* of B. Because B^n contains 2^n elements, this way one gets *all* finite Boolean algebras (out of isomorphism).

5.7.4 Boolean Algebras as Orderings

An order relation can be assigned to every Boolean algebra B: Here $a \le b$ holds if $a \sqcap b = a$ is valid (or equivalently, if $a \sqcup b = b$ holds).
So every finite Boolean algebra can be represented by a Hasse diagram (see 5.2.4, **4.**, p. 334).

■ Suppose B is the set $\{1, 2, 3, 5, 6, 10, 15, 30\}$ of the divisors of 30. Then, the least common multiple and the greatest common divisor can be defined as binary operations and the complement as unary operation. The numbers 1 and 30 correspond to the distinguished elements 0 and 1. The corresponding Hasse diagram is shown in **Fig. 5.20**.

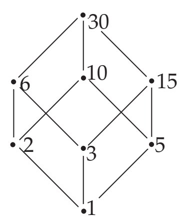


Figure 5.20

5.7.5 Boolean Functions, Boolean Expressions

1. Boolean Functions

Denoting by B the Boolean algebra with two elements as in 5.7.3, p. 397, then an *n-ary Boolean function* f is a mapping from B^n into B. There are 2^{2^n} n-ary Boolean functions. The set of all n-ary Boolean functions with the operations

$$(f \sqcap g)(b) = f(b) \sqcap g(b), \tag{5.304}$$

$$(f \sqcup g)(b) = f(b) \sqcup g(b), \tag{5.305}$$

$$\overline{f}(b) = \overline{f(b)}, \tag{5.306}$$

is a Boolean algebra. Here b always means an n tuple of the elements of $B = \{0, 1\}$, and on the right-hand side of the equations the operations are performed in B. The distinguished elements 0 and 1 correspond to the functions f_0 and f_1 with

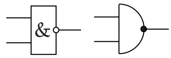
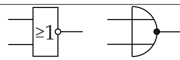
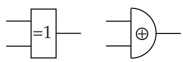
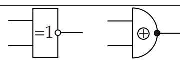
$$f_0(b) = 0, \quad f_1(b) = 1 \quad \text{for all} \quad b \in B^n. \tag{5.307}$$

■ **A:** In the case $n = 1$, i.e., for only one Boolean variable b, there are four Boolean functions:

$$\begin{array}{llll} \text{Identity} & f(b) = b, & \text{Negation} & f(b) = \overline{b}, \\ \text{Tautology} & f(b) = 1, & \text{Contradiction} & f(b) = 0. \end{array} \tag{5.308}$$

■ **B:** In the case $n = 2$, i.e., for two Boolean variables a and b, there are 16 different Boolean functions, among which the most important ones have their own names and notation. They are shown in **Table 5.6**.

Table 5.6 Some Boolean functions with two variables a and b

Name of the function	**Different notation**	**Different symbols**	**Value table for** $\begin{pmatrix}a\\b\end{pmatrix}=\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}0\\1\end{pmatrix},\begin{pmatrix}1\\0\end{pmatrix},\begin{pmatrix}1\\1\end{pmatrix}$
Sheffer or NAND	$\overline{a \cdot b}$ $a \mid b$ NAND (a, b)		1, 1, 1, 0
Peirce or NOR	$\overline{a + b}$ $a \downarrow b$ NOR a, b		1, 0, 0, 0
Antivalence or XOR	$\overline{a}\, b + a\, \overline{b}$ $a\,\mathrm{XOR}\,b$ $a \not\equiv b$ $a \oplus b$		0, 1, 1, 0
Equivalence	$\overline{a}\,\overline{b} + a\,b$ $a \equiv b$ $a \leftrightarrow b$		1, 0, 0, 1
Implication	$\overline{a} + b$ $a \to b$		1, 1, 0, 1

2. Boolean Expressions

Boolean expressions are defined in an inductive way: Let $X = \{x, y, z, \ldots\}$ be a (countable) set of *Boolean variables* (which can take values only from $\{0, 1\}$):

1. The constants 0 and 1 just as the Boolean variables from X are Boolean expressions. (5.309)

2. If S and T are Boolean expressions, so are $\overline{T}$, $(S \sqcap T)$, and $(S \sqcup T)$, as well. (5.310)

If a Boolean expression contains the variables $x_1, \ldots, x_n$, then it represents an n-ary Boolean function f_T:
Let b be a "valuation" of the Boolean variables $x_1, \ldots, x_n$, i.e., $b = (b_1, \ldots, b_n) \in B^n$.
Assigning a Boolean function to the expression T in the following way gives:

1. If $T = 0$, then $f_T = f_0$; if $T = 1$, then $f_T = f_1$. (5.311a)

2. If $T = x_i$, then $f_T(b) = b_i$; if $T = \overline{S}$, then $f_T(b) = \overline{f_S(b)}$. (5.311b)

3. If $T = R \sqcap S$, then $f_T(b) = f_R(b) \sqcap f_S(b)$. (5.311c)

4. If $T = R \sqcup S$, then $f_T(b) = f_R(b) \sqcup f_S(b)$. (5.311d)

On the other hand, every Boolean function f can be represented by a Boolean expression T (see 5.7.6, p. 399).

3. Concurrent or Semantically Equivalent Boolean Expressions

The Boolean expressions S and T are called *concurrent* or *semantically equivalent* if they represent the same Boolean function. Boolean expressions are equal if and only if they can be transformed into each other according to the axioms of a Boolean algebra.
Under transformations of a Boolean expression here are considered especially two aspects:

- Transformation in a possible "simple" form (see 5.7.7, p. 399).
- Transformation in a "normal form".

5.7.6 Normal Forms

1. Elementary Conjunction, Elementary Disjunction

Let $B = (B, \sqcap, \sqcup, ^{-}, 0, 1)$ be a Boolean algebra and $\{x_1, \ldots, x_n\}$ a set of Boolean variables. Every conjunction or disjunction in which every variable or its negation occurs exactly once is called an *elementary conjunction* or an *elementary disjunction* respectively (in the variables $x_1, \ldots, x_n$).
Let $T(x_1, \ldots, x_n)$ be a Boolean expression. A disjunction D of elementary conjunctions with $D = T$ is called a *principal disjunctive normal form* (*PDNF*) of T. A conjunction C of elementary disjunctions with $C = T$ is called a *principal conjunctive normal form* (*PCNF*) of T.

■ **Part 1:** In order to show that every Boolean function f can be represented as a Boolean expression, the PDNF form of the function f given in the annexed table is to be constructed:

x	y	z	$f(x,y,z)$
0	0	0	0
0	0	1	1
0	1	0	0
0	1	1	0
1	0	0	0
1	0	1	1
1	1	0	1
1	1	1	0

The PDNF of the Boolean function f contains the elementary conjunctions $\overline{x} \sqcap \overline{y} \sqcap z$, $x \sqcap \overline{y} \sqcap z$, $x \sqcap y \sqcap \overline{z}$. These elementary conjunctions belong to the valuations b of the variables where the function f has the value 1. If a variable v has the value 1 in b, then v is to put in the elementary conjunction, otherwise $\overline{v}$.

■ **Part 2:** The PDNF for the example of Part 1 is:

$$(\overline{x} \sqcap \overline{y} \sqcap z) \sqcup (x \sqcap \overline{y} \sqcap z) \sqcup (x \sqcap y \sqcap \overline{z}). \tag{5.312}$$

The "dual" form for PDNF is the PCNF: The elementary disjunctions belong to the valuations b of the variables for which f has the value 0. If a variable v has the value 0 in b, then v is to put in the elementary disjunction, otherwise $\overline{v}$. So the PCNF is:

$$(x \sqcup y \sqcup z) \sqcap (x \sqcup \overline{y} \sqcup z) \sqcap (x \sqcup \overline{y} \sqcup \overline{z}) \sqcap (\overline{x} \sqcup y \sqcup z) \sqcap (\overline{x} \sqcup \overline{y} \sqcup \overline{z}). \tag{5.313}$$

The PDNF and the PCNF of f are uniquely determined, if the ordering of the variables and the ordering of the valuations is given, e.g., if considering the valuations as binary numbers and arranging them in increasing order.

2. Principal Normal Forms

The principal normal form of a Boolean function f_T is considered as the principal normal form of the corresponding Boolean expression T.
Checking the equivalence of two Boolean expressions by transformations is often difficult. The principal normal forms are useful: Two Boolean expressions are semantically equivalent exactly if their corresponding uniquely determined principal normal forms are identical letter by letter.

■ **Part 3:** In the considered example (see Part 1 and 2) the expressions $(\overline{y} \sqcap z) \sqcup (x \sqcap y \sqcap \overline{z})$ and $(x \sqcup ((y \sqcup z) \sqcap (\overline{y} \sqcup z) \sqcap (\overline{y} \sqcup \overline{z}))) \sqcap (\overline{x} \sqcup ((y \sqcup z) \sqcap (\overline{y} \sqcup z)))$ are semantically equivalent because the principal disjunctive (or conjunctive) normal forms of both are the same.

5.7.7 Switch Algebra

A typical application of Boolean algebra is the simplification of series–parallel connections (SPC). Therefore a Boolean expression is to be assigned to a SPC (transformation). This expression will be "simplified" with the transformation rules of the Boolean algebra. Finally a SPC is to be assigned to this expression (inverse transformation). The result is a simplified SPC which produces the same behavior as the initial connection system **(Fig. 5.21)**.

A SPC has two types of contact points: the so-called "make contacts" and "break contacts", and both types have two states; namely open or closed. The usual symbolism is: When the equipment is put on, the make contacts close and the break contacts open. With Boolean variables assigned to the contacts of the switch equipment follows:
The position "off" or "on" of the equipment corresponds to the value 0 or 1 of the Boolean variables. The contacts being switched by the same equipment are denoted by the same symbol, the Boolean variable belonging to this equipment. The *contact value* of a SPC is 0 or 1, according to whether the switch is electrically non-conducting or conducting. The contact value depends on the position of the

contacts, so it is a Boolean function S (*switch function*) of the variables assigned to the switch equipment. Contacts, connections, symbols, and the corresponding Boolean expressions are represented in **Fig. 5.22**.

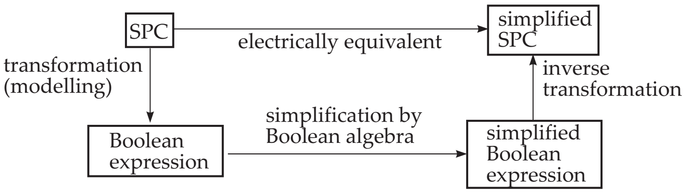


Figure 5.21

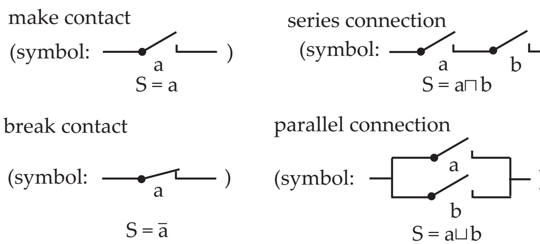


Figure 5.22

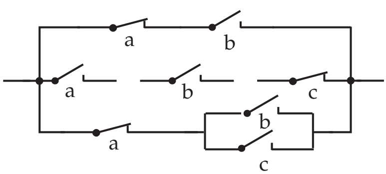


Figure 5.23

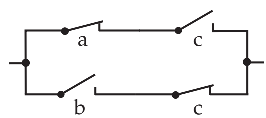


Figure 5.24

The Boolean expressions, which represent switch functions of SPC, have the special property that the negation sign can occur *only* above variables (never over subexpressions).

■ Simplification of the SPC **Fig. 5.23**. This connection corresponds to the Boolean expression

$$S = (\overline{a} \sqcap b) \sqcup (a \sqcap b \sqcap \overline{c}) \sqcup (\overline{a} \sqcap (b \sqcup c)) \tag{5.314}$$

as switch function. According to the transformation formulas of Boolean algebra holds:

$$\begin{aligned} S &= (b \sqcap (\overline{a} \sqcup (a \sqcap \overline{c}))) \sqcup (\overline{a} \sqcap (b \sqcup c)) \\ &= (b \sqcap (\overline{a} \sqcup \overline{c})) \sqcup (\overline{a} \sqcap (b \sqcup c)) \\ &= (\overline{a} \sqcap b) \sqcup (b \sqcap \overline{c}) \sqcup (\overline{a} \sqcap c) \\ &= (\overline{a} \sqcap b \sqcap c) \sqcup (\overline{a} \sqcap b \sqcap \overline{c}) \sqcup (b \sqcap \overline{c}) \sqcup (a \sqcap b \sqcap \overline{c}) \sqcup (\overline{a} \sqcap c) \sqcup (\overline{a} \sqcap \overline{b} \sqcap c) \\ &= (\overline{a} \sqcap c) \sqcup (b \sqcap \overline{c}). \end{aligned} \tag{5.315}$$

Here one gets $\overline{a} \sqcap c$ from $(\overline{a} \sqcap b \sqcap c) \sqcup (\overline{a} \sqcap c) \sqcup (\overline{a} \sqcap \overline{b} \sqcap c)$, and $b \sqcap \overline{c}$ from $(\overline{a} \sqcap b \sqcap \overline{c}) \sqcup (b \sqcap \overline{c}) \sqcup (a \sqcap b \sqcap \overline{c})$. The finally simplified result SPC is shown in **Fig. 5.24**.

This example shows that usually it is not so easy to get the simplest Boolean expression by transformations. In the literature one can find different methods for this procedure.

5.8 Algorithms of Graph Theory

Graph theory is a field in discrete mathematics having special importance for informatics, e.g., for representing data structures, finite automata, communication networks, derivatives in formal languages, etc. There are also applications in physics, chemistry, electrotechnics, biology and psychology. Moreover, flows can be applied in transport networks and in network analysis in operations research and in combinatorial optimization.

5.8.1 Basic Notions and Notation

1. Undirected and Directed Graphs

A *graph* G is an ordered pair (V, E) of a set V of *vertices* and a set E of *edges*. There is a mapping, defined on E, the *incidence function*, which uniquely assigns to every element of E an ordered or non-ordered pair of (not necessarily distinct) elements of V. If a non-ordered pair is assigned then G is called an *undirected graph* **(Fig. 5.25)**. If an ordered pair is assigned to every element of E, then the graph is called a *directed graph* **(Fig. 5.26)**, and the elements of E are called *arcs* or *directed edges*. All other graphs are called *mixed graphs*.
In the graphical representation, the vertices of a graph are denoted by points, the directed edges by arrows, and undirected edges by non-directed lines.

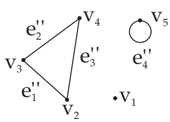


Figure 5.25

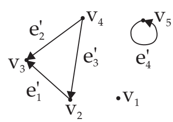


Figure 5.26

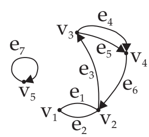


Figure 5.27

■ **A:** For the graph G in **Fig. 5.27**: $V = \{v_1, v_2, v_3, v_4, v_5\},\ E = \{e_1, e_2, e_3, e_4, e_5, e_6, e_7\}$,
$f_1(e_1) = \{v_1, v_2\},\ f_1(e_2) = \{v_1, v_2\},\ f_1(e_3) = (v_2, v_3),\ f_1(e_4) = (v_3, v_4),\ f_1(e_5) = (v_3, v_4)$,
$f_1(e_6) = (v_4, v_2),\ f_1(e_7) = (v_5, v_5)$.

■ **B:** For the graph G in **Fig. 5.26**: $V = \{v_1, v_2, v_3, v_4, v_5\}, \quad E' = \{e'_1, e'_2, e'_3, e'_4\}$
$f_2(e'_1) = (v_2, v_3),\ f_2(e'_2) = (v_4, v_3),\ f_2(e'_3) = (v_4, v_2),\ f_2(e'_4) = (v_5, v_5)$.

■ **C:** For the graph G in **Fig. 5.25**: $V = \{v_1, v_2, v_3, v_4, v_5\}, \quad E'' = \{e''_1, e''_2, e''_3, e''_4\}$,
$f_3(e''_1) = \{v_2, v_3\},\ f_3(e''_2) = \{v_4, v_3\},\ f_3(e''_3) = \{v_4, v_2\},\ f_3(e''_4) = \{v_5, v_5\}$.

2. Adjacency

If $(v, w) \in E$, then the vertex v is called *adjacent* to the vertex w. Vertex v is called the *initial point* of (v, w), w is called the *terminal point* of (v, w), and v and w are called the *endpoints* of (v, w).
Adjacency in undirected graphs and the endpoints of undirected edges are defined analogously.

3. Simple Graphs

If several edges or arcs are assigned to the same ordered or non-ordered pairs of vertices, then they are called *multiple edges*. An edge with identical endpoints is called a *loop*. Graphs without loops and multiple edges and multiple arcs, respectively, are called *simple* graphs.

4. Degrees of Vertices

The number of edges or arcs incident to a vertex v is called the *degree* $d_G(v)$ of the vertex v. Loops are counted twice. Vertices of degree zero are called *isolated vertices*.
For every vertex v of a directed graph G, the *out-degree* $d_G^+(v)$ and *in-degree* $d_G^-(v)$ of v are distinguished as follows:

$$d_G^+(v) = |\{w|(v, w) \in E\}|, \qquad (5.316a) \qquad\qquad d_G^-(v) = |\{w|(w, v) \in E\}|. \qquad (5.316b)$$

5. Special Classes of Graphs

Finite graphs have a finite set of vertices and a finite set of edges. Otherwise the graph is called *infinite*.
In *regular graphs of degree* r every vertex has degree r.
An undirected simple graph with vertex set V is called a *complete graph* if any two different vertices in V are connected by an edge. A complete graph with an n element set of vertices is denoted by K_n.

If the set of vertices of an undirected simple graph G can be partitioned into two disjoint classes X and Y such that every edge of G joins a vertex of X and a vertex of Y, then G is called a *bipartite graph*.
A bipartite graph is called a *complete bipartite graph*, if every vertex of X is joined by an edge with every vertex of Y. If X has n elements and Y has m elements, then the graph is denoted by $K_{n,m}$.

■ **Fig. 5.28** shows a complete graph with five vertices.

■ **Fig. 5.29** shows a complete bipartite graph with a two-element set X and a three-element set Y.

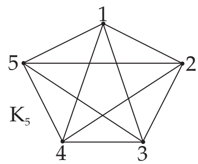


Figure 5.28

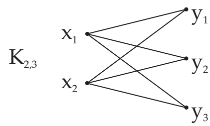


Figure 5.29

Further special classes of graphs are *plane graphs*, *trees* and *transport networks*. Their properties will be discussed in later paragraphs.

6. Representation of Graphs

Finite graphs can be visualized by assigning to every vertex a point in the plane and connecting two points by a directed or undirected curve, if the graph has the corresponding edge. There are examples in **Fig. 5.30**–**5.33**. **Fig. 5.33** shows the *Petersen graph*, which is a well-known counterexample for several graph-theoretic conjectures, which could not be proved in general.

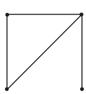


Figure 5.30

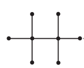


Figure 5.31

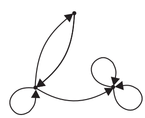


Figure 5.32

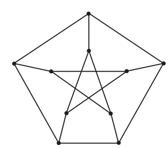


Figure 5.33

7. Isomorphism of Graphs

A graph $G_1 = (V_1, E_1)$ is called *isomorphic* to a graph $G_2 = (V_2, E_2)$ iff there are bijective mappings φ from V_1 onto V_2 and ψ from E_1 onto E_2 being compatible with the incidence function, i.e., if u, v are the endpoints of an edge or u is the initial point of an arc and v is its terminal point, then $\varphi(u)$ and $\varphi(v)$ are the endpoints of an edge and $\varphi(u)$ is the initial point and $\varphi(v)$ the terminal point of an arc, respectively. **Fig. 5.34** and **Fig. 5.35** show two isomorphic graphs. The mapping φ with $\varphi(1) = a$, $\varphi(2) = b$, $\varphi(3) = c$, $\varphi(4) = d$ is an isomorphism. In this case, every bijective mapping of $\{1, 2, 3, 4\}$ onto $\{a, b, c, d\}$ is an isomorphism, since both graphs are complete graphs with equal number of vertices.

8. Subgraphs, Factors

If $G = (V, E)$ is a graph, then the graph $G' = (V', E')$ is called a *subgraph* of G, if $V' \subseteq V$ and $E' \subseteq E$.
If E' contains exactly those edges of E which connect vertices of V', then G' is called the *subgraph of* G *induced by* V' (*induced subgraph*).

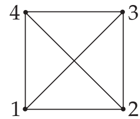


Figure 5.34

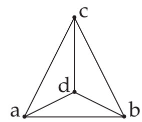


Figure 5.35

A subgraph $G' = (V', E')$ of $G = (V, E)$ with $V' = V$ is called a *partial graph* of G.
A factor F of a graph G is a regular subgraph of G containing all vertices of G.

9. Adjacency Matrix

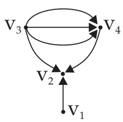
Finite graphs can be described by matrices: Let $G = (V, E)$ be a graph with $V = \{v_1, v_2, \ldots, v_n\}$ and $E = \{e_1, e_2, \ldots, e_m\}$. Let $m(v_i, v_j)$ denote the number of edges from v_i to v_j. For undirected graphs, loops are counted twice; for directed graphs loops are counted once. The matrix $\mathbf{A}$ of type (n, n) with $\mathbf{A} = (m(v_i, v_j))$ is called an *adjacency matrix*. If in addition the graph is simple, then the adjacency matrix has the following form:

$$\mathbf{A} = (a_{ij}) = \begin{cases} 1, & \text{for } (v_i, v_j) \in E, \\ 0, & \text{for } (v_i, v_j) \notin E; \end{cases} \tag{5.317}$$

i.e., in the matrix $\mathbf{A}$ there is a 1 in the i-th row and j-th column iff there is an edge from v_i to v_j.
The adjacency matrix of undirected graphs is symmetric.

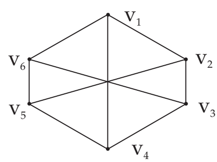
■ **A:** Beside **Fig. 5.36** there is the adjacency matrix $\mathbf{A}_1 = \mathbf{A}(G_1)$ of the directed graph G_1.

■ **B:** Beside **Fig. 5.37** there is the adjacency matrix $\mathbf{A}_2 = \mathbf{A}(G_2)$ of the undirected simple graph G_2.



$$\mathbf{A}_1 = \begin{pmatrix} 0&1&0&0 \\ 0&0&0&0 \\ 0&1&0&3 \\ 0&1&0&0 \end{pmatrix}$$

Figure 5.36



$$\mathbf{A}_2 = \begin{pmatrix} 0&1&0&1&0&1 \\ 1&0&1&0&1&0 \\ 0&1&0&1&0&1 \\ 1&0&1&0&1&0 \\ 0&1&0&1&0&1 \\ 1&0&1&0&1&0 \end{pmatrix}$$

Figure 5.37

10. Incidence Matrix

For an undirected graph $G = (V, E)$ with $V = \{v_1, v_2, \ldots, v_n\}$ and $E = \{e_1, e_2, \ldots, e_m\}$, the matrix $\mathbf{I}$ of type (n, m) given by

$$\mathbf{I} = (b_{ij}) \text{ with } b_{ij} = \begin{cases} 0, & v_i \text{ is not incident with } e_j, \\ 1, & v_i \text{ is incident with } e_j \text{ and } e_j \text{ is not a loop}, \\ 2, & v_i \text{ is incident with } e_j \text{ and } e_j \text{ is a loop} \end{cases} \tag{5.318}$$

is called the *incidence matrix*.
For a directed graph $G = (V, E)$ with $V = \{v_1, v_2, \ldots, v_n\}$ and $E = \{e_1, e_2, \ldots, e_m\}$, the incidence matrix $\mathbf{I}$ is the matrix of type (n, m), defined by

$$\mathbf{I} = (b_{ij}) \text{ with } b_{ij} = \begin{cases} 0, & v_i \text{ is not incident with } e_j, \\ 1, & v_i \text{ is the initial point of } e_j \text{ and } e_j \text{ is not a loop}, \\ -1, & v_i \text{ is the terminal point of } e_j \text{ and } e_j \text{ is not a loop}, \\ -0, & v_i \text{ is incident to } e_j \text{ and } e_j \text{ is a loop}. \end{cases} \tag{5.319}$$

11. Weighted Graphs

If $G = (V, E)$ is a graph and f is a mapping assigning a real number to every edge, then (V, E, f) is called a *weighted graph*, and $f(e)$ is the *weight* or *length* of the edge e.

In applications, these weights of the edges represent costs resulting from the construction, maintenance or use of the connections.

5.8.2 Traverse of Undirected Graphs

5.8.2.1 Edge Sequences or Paths

1. Edge Sequences or Paths

In an undirected graph $G = (V, E)$ every sequence $F = (\{v_1, v_2\}, \{v_2, v_3\}, \dots, \{v_s, v_{s+1}\})$ of the elements of E is called an *edge sequence* of length s.
If $v_1 = v_{s+1}$, then the sequence is called a *cycle*, otherwise it is an *open edge sequence*. An edge sequence F is called a *path* iff $v_1, v_2, \dots, v_s$ are pairwise distinct vertices. A *closed path* is a *circuit*. A *trail* is a sequence of edges without repeated edges.

■ In the graphs in **Fig. 5.38**, $F_1 = (\{1,2\}, \{2,3\}, \{3,5\}, \{5,2\}, \{2,4\})$ is an edge sequence of length 5, $F_2 = (\{1,2\}, \{2,3\}, \{3,4\}, \{4,2\}, \{2,1\})$ is a cycle of length 5, $F_3 = (\{2,3\}, \{3,5\}, \{5,2\}, \{2,1\})$ is a path, $F_4 = (\{1,2\}, \{2,3\}, \{3,4\})$ is a path. An elementary cycle is given by $F_5 = (\{1,2\}, \{2,5\}, \{5,1\})$.

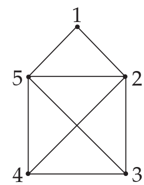


Figure 5.38

2. Connected Graphs, Components

If there is at least one path between every pair of distinct vertices v, w in a graph G, then G is called *connected*. If a graph G is not connected, it can be decomposed into *components*, i.e., into induced connected subgraphs with maximal number of vertices.

3. Distance Between Vertices

The distance $\delta(v, w)$ between two vertices v, w of an undirected graph is the length of a path with minimum number of edges connecting v and w. If such a path does not exist, then let $\delta(v, w) = \infty$.

4. Problem of Shortest Paths

Let $G = (V, E, f)$ be a weighted simple graph with $f(e) > 0$ for every $e \in E$. Determine the *shortest path* from v to w for two vertices v, w of G, i.e., a path from v to w having minimum sum of weights of edges and arcs, respectively.
There is an efficient algorithm of Dantzig to solve this problem, which is formulated for directed graphs and can be used for undirected graphs (see 5.8.6, p. 410) in a similar way.
Every graph $G = (V, E, f)$ with $V = \{v_1, v_2, \dots, v_n\}$ has a *distance matrix* $\mathbf{D}$ of type (n, n):

$$\mathbf{D} = (d_{ij}) \quad \text{with} \quad d_{ij} = \delta(v_i, v_j) \qquad (i, j = 1, 2, \dots, n). \tag{5.320}$$

In the case that every edge has weight 1, i.e., the distance between v and w is equal to the minimum number of edges which have to be traversed in the graph to get from v to w, then the distance between two vertices can be determined using the adjacency matrix: Let $v_1, v_2, \dots, v_n$ be the vertices of G. The adjacency matrix of G is $\mathbf{A} = (a_{ij})$, and the powers of the adjacency matrix with respect to the usual multiplication of matrices (see 4.1.4, **5.**, p. 272) are denoted by $\mathbf{A}^m = (a_{ij}^m)$, $m \in \mathbb{N}$.
There is a shortest path of length k from the vertex v_i to the vertex v_j $(i \neq j)$ iff:

$$a_{ij}^k \neq 0 \quad \text{and} \quad a_{ij}^s = 0 \quad (s = 1, 2, \dots, k-1). \tag{5.321}$$

■ The weighted graph represented in **Fig. 5.39** has the distance matrix $\mathbf{D}$ beside it.

■ The graph represented in **Fig. 5.40** has the adjacency matrix $\mathbf{A}$ beside it, and for $m = 2$ or $m = 3$ the matrices $\mathbf{A}^2$ and $\mathbf{A}^3$ are obtained. Shortest paths of length 2 connect the vertices 1 and 3, 1 and 4, 1 and 5, 2 and 6, 3 and 4, 3 and 5, 4 and 5. Furthermore the shortest paths between the vertices 1 and 6, 3 and 6, and finally 4 and 6 are of length 3.

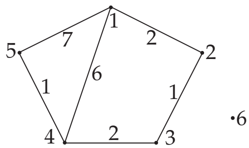


Figure 5.39

$$\mathbf{D} = \begin{pmatrix} 0 & 2 & 3 & 5 & 6 & \infty \\ 2 & 0 & 1 & 3 & 4 & \infty \\ 3 & 1 & 0 & 2 & 3 & \infty \\ 5 & 3 & 2 & 0 & 1 & \infty \\ 6 & 4 & 3 & 1 & 0 & \infty \\ \infty & \infty & \infty & \infty & \infty & 0 \end{pmatrix}$$

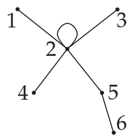


Figure 5.40

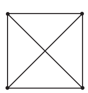
$$\mathbf{A} = \begin{pmatrix} 0&1&0&0&0&0 \\ 1&1&1&1&1&0 \\ 0&1&0&0&0&0 \\ 0&1&0&0&0&0 \\ 0&1&0&0&0&1 \\ 0&0&0&0&1&0 \end{pmatrix}, \quad \mathbf{A}^2 = \begin{pmatrix} 1&1&1&1&1&0 \\ 1&5&1&1&1&1 \\ 1&1&1&1&1&0 \\ 1&1&1&1&1&0 \\ 1&1&1&1&2&0 \\ 0&1&0&0&0&1 \end{pmatrix}, \quad \mathbf{A}^3 = \begin{pmatrix} 1&5&1&1&1&1 \\ 5&9&5&5&6&1 \\ 1&5&1&1&1&1 \\ 1&5&1&1&1&1 \\ 1&6&1&1&1&2 \\ 1&1&1&1&2&0 \end{pmatrix}.$$

5.8.2.2 Euler Trails

1. Euler Trail, Euler Graph

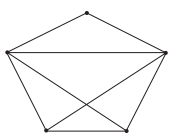
A trail containing every edge of a graph G is called an *open* or *closed Euler trail* of G.
A connected graph containing a closed Euler trail is an *Euler graph.*

■ The graph G_1 **(Fig. 5.41)** has no Euler trail. The graph G_2 **(Fig. 5.42)** has an Euler trail, but it is not an Euler graph. The graph G_3 **(Fig. 5.43)** has a closed Euler trail, but it is not an Euler graph. The graph G_4 **(Fig. 5.44)** is an Euler graph.



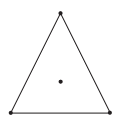
G_1

Figure 5.41



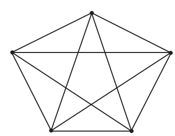
G_2

Figure 5.42



G_3

Figure 5.43



G_4

Figure 5.44

2. Theorem of Euler-Hierholzer

A finite connected graph is an Euler graph iff all vertices have positive even degrees.

3. Construction of a Closed Euler Trail

If G is an Euler graph, then one chooses an arbitrary vertex v_1 of G and constructs a trail F_1 by traversing a path, starting at v_1 and proceeding until it cannot be continued. If F_1 does not yet contain all edges of G, then one constructs another path F_2 containing the edges not in F_1, but starting at a vertex $v_2 \in F_1$ and proceeds until it cannot be continued. Then one composes a closed trail in G using F_1 and F_2: Starting to traverse F_1 at v_1 until v_2 is reached, then continuing to traverse F_2, and finishing at the edges of F_1 not used before. Repeating this method a closed Euler trail is obtained in finitely many steps.

4. Open Euler Trails

There is an open Euler trail in a graph G iff there are exactly two vertices in G with odd degrees. **Fig. 5.45** shows a graph which has no closed Euler trail, but it has an open Euler trail. The edges are consecutively enumerated with respect to an Euler trail. In **Fig. 5.46** there is a graph with a closed Euler trail.

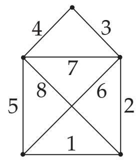


Figure 5.45

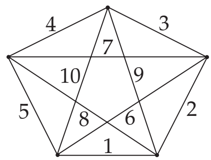


Figure 5.46

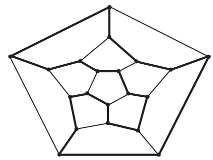


Figure 5.47

5. Chinese Postman Problem

The problem, that a postman should pass through all streets in his service area at least once and return to the initial point and use a trail as short as possible, can be formulated in graph theoretical terms as follows: Let $G = (V, E, f)$ be a weighted graph with $f(e) \geq 0$ for every edge $e \in E$. Determine an edge sequence F with minimum total length

$$L = \sum_{e \in F} f(e). \tag{5.322}$$

The name of the problem refers to the Chinese mathematician Kuan, who studied this problem first. To solve it two cases are distinguished:

1. G is an Euler graph – then every closed Euler trail is optimal – and

2. G has no closed Euler trail.

An effective algorithm solving this problem is given by Edmonds and Johnson (see [5.25]).

5.8.2.3 Hamiltonian Cycles

1. Hamiltonian Cycle

A *Hamiltonian cycle* is an elementary cycle in a graph covering all of the vertices.

■ In **Fig. 5.47**, lines in bold face show a Hamiltonian cycle.

The idea of a game to construct Hamiltonian cycles in the graph of a pentagondodecaeder, goes back to Sir W. Hamilton.

Remark: The problem of characterizing graphs with Hamiltonian cycles leads to one of the classical NP-complete problems. Therefore, an efficient algorithm to determine the Hamilton cycles cannot be given here.

2. Theorem of Dirac

If a simple graph $G = (V, E)$ has at least three vertices, and $d_G(v) \geq |V|/2$ holds for every vertex v of G, then G has a Hamiltonian cycle. This is a sufficient but not a necessary condition for the existence of Hamiltonian cycles. The following theorems with more general assumptions give only sufficient but not necessary conditions for the existence of Hamilton cycles, too.

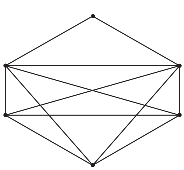


Figure 5.48

■ **Fig. 5.48** shows a graph which has a Hamiltonian cycle, but does not satisfy the assumptions of the following theorem of Ore.

3. Theorem of Ore

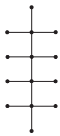
If a simple graph $G = (V, E)$ has at least three vertices, and $d_G(v) + d_G(w) \geq |V|$ holds for every pair of non-adjacent vertices v, w, then G contains a Hamiltonian cycle.

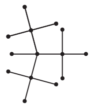
4. Theorem of Posa

Let $G = (V, E)$ be a simple graph with at least three vertices. There is a Hamiltonian cycle in G if the following conditions are satisfied:

1. For $1 \leq k < (|V| - 1)/2$, the number of vertices of degree not exceeding k is less than k.

2. If $|V|$ is odd, then the number of vertices of degree not exceeding $(|V| - 1)/2$ is less than or equal to $(|V| - 1)/2$.


Figure 5.49


Figure 5.50

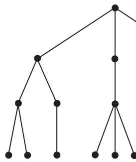


Figure 5.51

5.8.3 Trees and Spanning Trees

5.8.3.1 Trees

1. Trees

An undirected connected graph without cycles is called a *tree*. Every tree with at least two vertices has at least two vertices of degree 1. Every tree with n vertices has exactly $n - 1$ edges.
A directed graph is called a tree if G is connected and does not contain any circuit (see 5.8.6, p. 410).

■ **Fig. 5.49** and **Fig. 5.50** represent two non-isomorphic trees with 14 vertices. They demonstrate the chemical structure of butane and iso-butane.

2. Rooted Trees

A tree with a distinguished vertex is called a rooted tree, and the distinguished vertex is called the *root*. In diagrams, the root is usually on the top, and the edges are directed downwards from the root (see **Fig. 5.51**). Rooted trees are used to represent hierarchic structures, as for instance hierarchies in factories, family trees, grammatical structures.

■ **Fig. 5.51** shows the genealogy of a family in the form of a rooted tree. The root is the vertex assigned to the father.

3. Regular Binary Trees

If a tree has exactly one vertex of degree 2 and otherwise only vertices of degree 1 or 3, then it is called a *regular binary tree*.
The number of vertices of a regular binary tree is odd. Regular trees with n vertices have $(n + 1)/2$ vertices of degree 1. The *level* of a vertex is its distance from the root. The maximal level occurring in a tree is the *height* of the tree. There are several applications of regular binary rooted trees, e.g., in informatics.

4. Ordered Binary Trees

Arithmetical expressions can be represented by binary trees. Here, the numbers and variables are assigned vertices of degree 1, the operations "+","−", "·" correspond to vertices of degree > 1, and the left and right subtree, respectively, represents the first and second operand, respectively, which is, in general, also an expression. These trees are called *ordered binary trees*.
The traverse of an ordered binary tree can be performed in three different ways, which are defined in a recursive way (see also **Fig. 5.52**):

Inorder traverse: Traverse the left subtree of the root (in inorder traverse),
visit the root,
traverse the right subtree of the root (in inorder traverse).

Preorder traverse: Visit the root,
traverse the left subtree (in preorder traverse),
traverse the right subtree of the root (in preorder traverse).

Postorder traverse: Traverse the left subtree of the root (in postorder traverse),
traverse the right subtree of the root (in postorder traverse),
visit the root.

Using inorder traverse the order of the terms does not change in comparison with the given expression. The term obtained by postorder traverse is called *postfix notation* PN or *Polish notation*. Analogously, the term obtained by preorder traverse is called *prefix notation* or *reversed Polish notation*.

Prefix and postfix expressions uniquely describe the tree. This fact can be used for the implementation of trees.

■ In **Fig. 5.52** the term $a \cdot (b - c) + d$ is represented by a graph. Inorder traverse yields $a \cdot b - c + d$, preorder traverse yields $+ \cdot -bcad$, and postorder traversal yields $abc - \cdot d+$.

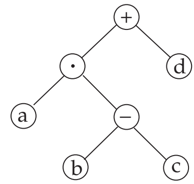


Figure 5.52

5.8.3.2 Spanning Trees

1. Spanning Trees

A tree, being a subgraph of an undirected graph G, and containing all vertices of G, is called a *spanning tree* of G. Every finite connected graph G contains a spanning tree H:

If G contains a cycle, then delete an edge of this cycle. The remaining graph G_1 is still connected and can be transformed into a connected graph G_2 by deleting a further edge of a cycle of G_1, if there exists such an edge. After finitely many steps a spanning tree of G is obtained.

■ **Fig. 5.54** shows a spanning tree H of the graph G shown in **Fig. 5.53**.

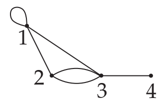


Figure 5.53

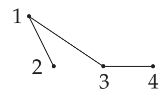


Figure 5.54

2. Theorem of Cayley

Every complete graph with n vertices ($n > 1$) has exactly n^{n-2} spanning trees.

3. Matrix Spanning Tree Theorem

Let $G = (V, E)$ be a graph with $V = \{v_1, v_2, \dots, v_n\}$ ($n > 1$) and $E = \{e_1, e_2, \dots, e_m\}$. Define a matrix $\mathbf{D} = (d_{ij})$ of type (n, n):

$$d_{ij} = \begin{cases} 0 \text{ for } i \neq j, \\ d_G(v_i) \text{ for } i = j, \end{cases} \tag{5.323a}$$

which is called the *degree matrix*. The difference between the degree matrix and the adjacency matrix is the admittance matrix $\mathbf{L}$ of G:

$$\mathbf{L} = \mathbf{D} - \mathbf{A}. \tag{5.323b}$$

Deleting the i-th row and the i-th column of $\mathbf{L}$ the matrix $\mathbf{L}_i$ is obtained. The determinant of $\mathbf{L}_i$ is equal to the number of spanning trees of the graph G.

■ The adjacency matrix, the degree matrix and the admittance matrix of the graph in **Fig. 5.53** are:

$$\mathbf{A} = \begin{pmatrix} 2&1&1&0 \\ 1&0&2&0 \\ 1&2&0&1 \\ 0&0&1&0 \end{pmatrix}, \qquad \mathbf{D} = \begin{pmatrix} 4&0&0&0 \\ 0&3&0&0 \\ 0&0&4&0 \\ 0&0&0&1 \end{pmatrix}, \qquad \mathbf{L} = \begin{pmatrix} 2&-1&-1&0 \\ -1&3&-2&0 \\ -1&-2&4&-1 \\ 0&0&-1&1 \end{pmatrix}.$$

Since $\det \mathbf{L}_3 = 5$, the graph has five spanning trees.

4. Minimal Spanning Trees

Let $G = (V, E, f)$ be a connected weighted graph. A spanning tree H of G is called a *minimum spanning tree* if its *total length* $f(H)$ is minimum:

$$f(H) = \sum_{e \in H} f(e). \tag{5.324}$$

Minimum spanning trees are searched for, e.g., if the edge weights represent costs, and one is interested in minimum costs. A method to find a minimum spanning tree is the *Kruskal algorithm*:

a) Choose an edge with the least weight.

b) Continue, as long as it is possible, choosing a further edge having least weight and not forming a cycle with the edges already chosen, and add such an edge to the tree.

In step **b)** the choice of the admissible edges can be made easier by the following labeling algorithm:
• Let the vertices of the graph be labeled pairwise differently.
• At every step, an edge can be added only in the case that it connects vertices with different labels.
• After adding an edge, the label of the endpoint with the larger label is changed to the value of the smaller endpoint label.

5.8.4 Matchings

1. Matchings

A set M of edges of a graph G is called a *matching* in G, iff M contains no loop and two different edges of M do not have common endpoints.
A matching M^* of G is called a *saturated matching*, if there is no matching M in G such that $M^* \subset M$.
A matching M^{**} of G is called a *maximum matching*, if there is no matching M in G such that $|M| > |M^{**}|$.
If M is a matching of G such that every vertex of G is an endpoint of an edge of M, then M is called a *perfect matching.*

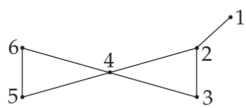


Figure 5.55

■ In the graph in **Fig. 5.55** $M_1 = \{\{2,3\},\{5,6\}\}$ is a saturated matching and $M_2 = \{\{1,2\},\{3,4\},\{5,6\}\}$ is a maximum matching which is also perfect.

Remark: In graphs with an odd number of edges there is no perfect matching.

2. Theorem of Tutte

Let $q(G - S)$ denote the number of the components of $G - S$ with an odd number of vertices. A graph $G = (V, E)$ has a perfect matching iff $|V|$ is even and for every subset S of the vertex set $q(G-S) \leq |S|$. Here $G - S$ denotes the graph obtained from G by deleting the vertices of S and the edges incident with these vertices.
Perfect machings exist for example in complete graphs with an even number of vertices, in complete bipartite graphs $K_{n,n}$ and in arbitrary regular bipartite graphs of degree $r > 0$.

3. Alternating Paths

Let G be a graph with a matching M. A path W in G is called an *alternating path* iff in W every edge e with $e \in M$ (or $e \notin M$) is followed by an edge e' with $e' \notin M$ (or $e \in M$).
An open alternating path is called an *increasing path* iff none of the endpoints of the path is incident with an edge of M.

4. Theorem of Berge

A matching M in a graph G is maximum iff there is no increasing alternating path in G.
If W is an increasing alternating path in G with corresponding set $E(W)$ of traversed edges, then $M' = (M \setminus E(W)) \cup (E(W) \setminus M)$ forms a matching in G with $|M'| = |M| + 1$.

■ In the graph of **Fig. 5.55** $(\{1,2\},\{2,3\},\{3,4\})$ is an increasing alternating path with respect to matching M_1. Matching M_2 with $|M_2| = |M_1| + 1$ is obtained as described above.

5. Determination of Maximum Matchings

Let G be a graph with a matching M.

a) First form a saturated matching M^* with $M \subseteq M^*$.

b) Chose a vertex v in G, which is not incident with an edge of M^*, and determine an increasing alternating path in G starting at v.

c) If such a path exists, then the method described above results in a matching M' with $|M'| > |M^*|$. If there is no such path, then delete vertex v and all edges incident with v in G, and repeat step **b)**.

There is an algorithm of Edmonds, which is an effective method to search for maximum matchings, but it is rather complicated to describe (see [5.24]).

5.8.5 Planar Graphs

Here, the considerations are restricted to undirected graphs, since a directed graph is planar iff the corresponding undirected graph is a planar one.

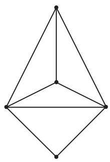


Figure 5.56

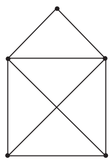


Figure 5.57

1. Planar Graph

A graph is called a *plane graph* iff G can be drawn in the plane with its edges intersecting only in vertices of G. A graph isomorphic with a plane graph is called a *planar graph*.

Fig. 5.56 shows a plane graph G_1. The graph G_2 in **Fig. 5.57** is isomorphic to G_1, it is not a plane graph but a planar graph, since it is isomorphic with G_1.

2. Non-Planar Graphs

The complete graph K_5 and the complete bipartite graph $K_{3,3}$ are non-planar graphs (see 5.8.1, **5.**, p. 402).

3. Subdivisions

A *subdivision* of a graph G is obtained if vertices of degree 2 are inserted into edges of G. Every graph is a subdivision of itself. Certain subdivisions of K_5 and $K_{3,3}$ are represented in **Fig. 5.58** and **Fig. 5.59**.

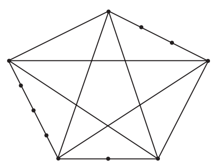


Figure 5.58

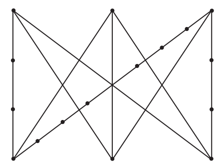


Figure 5.59

4. Kuratowski's Theorem

A graph is non-planar iff it contains a subgraph which is a subdivision either of the complete bipartite graph $K_{3,3}$ or of the complete graph K_5.

5.8.6 Paths in Directed Graphs

1. Arc Sequences

A sequence $F = (e_1, e_2, \ldots, e_s)$ of arcs in a directed graph is called a *chain* of length s, iff F does not contain any arc twice and one of the endpoints of every arc e_i for $i = 2, 3, \ldots, s-1$ is an endpoint of the arc e_{i-1} and the other one an endpoint of e_{i+1}.

A chain is called a *directed chain* iff for $i = 1, 2, \ldots, s-1$ the terminal point of the arc e_i coincides with the initial point of e_{i+1}.
Chains or directed chains traversing every vertex at most once are called *elementary chains* and *elementary directed chains*, respectively.
A closed chain is called a *cycle*. A closed directed path, with every vertex being the endpoint of exactly two arcs, is called a *circuit*.

■ **Fig. 5.60** contains examples for various kinds of arc sequences.

2. Connected and Strongly Connected Graphs

A directed graph G is called *connected* iff for any two vertices there is a chain connecting these vertices. The graph G is called *strongly connected* iff to every two vertices v, w there is is assigned a directed chain connecting these vertices.

3. Algorithm of Dantzig

Let $G = (V, E, f)$ be a weighted simple directed graph with $f(e) > 0$ for every arc e. The following algorithm yields all vertices of G, which are connected with a fixed vertex v_1 by a directed chain, together with their distances from v_1:

a) Vertex v_1 gets the label $t(v_1) = 0$. Let $S_1 = \{v_1\}$.

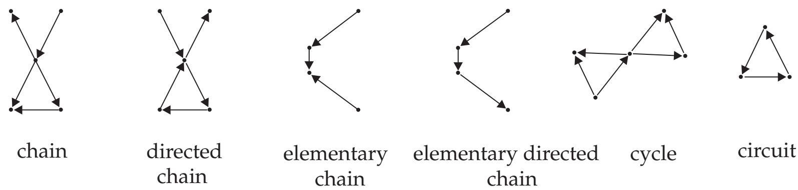


Figure 5.60

b) The set of the labeled vertices is S_m.

c) If $U_m = \{e|e = (v_i, v_j) \in E,\ v_i \in S_m,\ v_j \notin S_m\} = \emptyset$, then one finishes the algorithm.

d) Otherwise one chooses an arc $e^* = (x^*, y^*)$ with minimum $t(x^*) + f(e^*)$. One labels e^* and y^* and puts $t(y^*) = t(x^*) + f(e^*)$ and also $S_{m+1} = S_m \cup \{y^*\}$ and repeats **b)** with $m := m + 1$.

(If all arcs have weight 1, then the length of a shortest directed chain from a vertex v to a vertex w can be found using the adjacency matrix (see 5.8.2.1, **4.**, p. 404)).

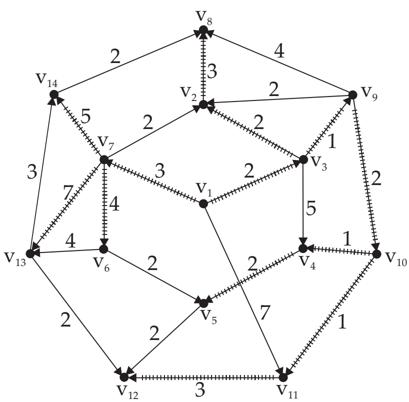


Figure 5.61

If a vertex v of G is not labeled, then there is no directed path from v_1 to v.

If v has label $t(v)$, then $t(v)$ is the length of such a directed chain. A shortest directed path from v_1 to v can be found in the tree given by the labeled arcs and vertices, the *distance tree* with respect to v_1.

■ In **Fig. 5.61**, the labeled arcs and vertices represent the distance tree with respect to v_1 in the graph. The lengths of the shortest directed chains are:

from v_1 to v_3 :	2	from v_1 to v_6 :	7
from v_1 to v_7 :	3	from v_1 to v_8 :	7
from v_1 to v_9 :	3	from v_1 to v_{14} :	8
from v_1 to v_2 :	4	from v_1 to v_5 :	8
from v_1 to v_{10} :	5	from v_1 to v_{12} :	9
from v_1 to v_4 :	6	from v_1 to v_{13} :	10
from v_1 to v_{11} :	6.		

Remark: There is also a modified algorithm to find the shortest directed chains in the case that $G = (V, E, f)$ has arcs with negative weights.

5.8.7 Transport Networks

1. Transport Network

A connected directed graph is called a *transport network* if it has two labeled vertices, called the *source* Q and *sink* S which have the following properties:

a) There is an arc u_1 from S to Q, where u_1 is the only arc with initial point S and the only arc with terminal point Q.

b) Every arc u_i different from u_1 is assigned a real number $c(u_i) \geq 0$. This number is called its *capacity*. The arc u_1 has capacity ∞.

A function φ, which assigns a real number to every arc, is called a *flow* on G, if the equality

$$\sum_{(u,v)\in G} \varphi(u, v) = \sum_{(v,w)\in G} \varphi(v, w) \tag{5.325a}$$

holds for every vertex v. The sum

$$\sum_{(Q,v)\in G} \varphi(Q,v) \tag{5.325b}$$

is called the intensity of the flow. A flow φ is called *compatible to the capacities*, if for every arc u_i of G $0 \le \varphi(u_i) \le c(u_i)$ holds.

■ For an example of a transport network see p. 412.

2. Maximum Flow Algorithm of Ford and Fulkerson

Using the maximum flow algorithm one can recognize whether a given flow φ is maximal.
Let G be a transport network and φ a flow of intensity v_1 compatible with the capacities. The algorithm given below contains the following steps for labeling the vertices, and after finishing this procedure one can realize how much the intensity of the flow could be improved depending on the chosen labeling steps.

a) One labels the source Q and sets $\varepsilon(Q) = \infty$.

b) If there is an arc $u_i = (x,y)$ with labeled x and unlabeled y and $\varphi(u_i) < c(u_i)$, then one labels y and (x,y), and sets $\varepsilon(y) = \min\{\varepsilon(x), c(u_i) - \varphi(u_i)\}$, then one repeats step **b)**, otherwise follows step **c)**.

c) If there is an arc $u_i = (x,y)$ with unlabeled x and labeled y, $\varphi(u_i) > 0$ and $u_i \ne u_1$, then one labels x and (x,y), substitutes $\varepsilon(x) = \min\{\varepsilon(y), \varphi(u_i)\}$ and returns to continue step **b)** if it is possible. Otherwise one finishs the algorithm.

If the sink S of G is labeled, then the flow in G can be improved by an amount of $\varepsilon(S)$. If the sink is not labeled, then the flow is maximal.

■ Maximum flow: For the graph in **Fig. 5.62** the weights are written next to the edges. A flow with intensity 13, compatible to these capacities, is represented in the weighted graph in **Fig. 5.63**. It is a maximum flow.

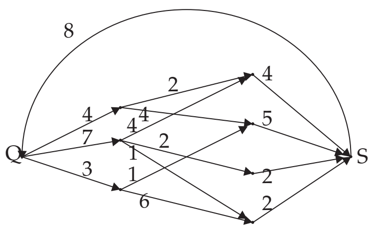


Figure 5.62

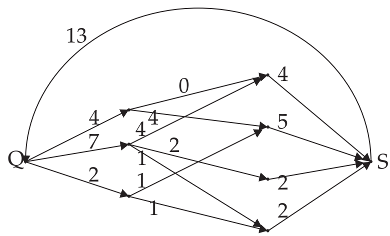


Figure 5.63

■ Transport network: A product is produced by p firms $F_1, F_2, \ldots, F_p$. There are q users $V_1, V_2, \ldots, V_q$. During a certain period there will be s_i units produced by F_i and t_j units required by V_j. c_{ij} units can be transported from F_i to V_j during the given period. Is it possible to satisfy all the requirements during this period? The corresponding graph is shown in **Fig. 5.64**.

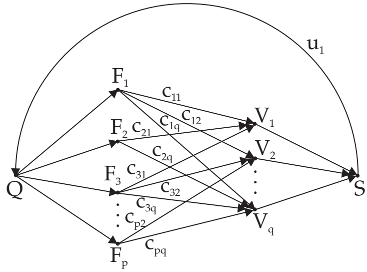


Figure 5.64

5.9 Fuzzy Logic

5.9.1 Basic Notions of Fuzzy Logic

5.9.1.1 Interpretation of Fuzzy Sets

Real situations are very often uncertain or vague in a number of ways. The word “fuzzy” also means some uncertainty, and the name of *fuzzy logic* is based on this meaning. Basically there are to distinguish two types of fuzziness: *vagueness* and *uncertainty*. There are two concepts belonging here: The theory of fuzzy sets and the theory of fuzzy measure. In the following practice-oriented introduction the notions, methods, and concepts of fuzzy sets are discussed, which are the basic mathematical tools of multi-valued logic.

1. Notions of Classical and Fuzzy Sets

The classical notion of (crisp) set is two-valued, and the classical Boolean set algebra is isomorphic to two-valued propositional logic. Let X be a fundamental set named the universe. Then for every $A \subseteq X$ there exists a function

$$f_A\colon X \to \{0,1\}, \tag{5.326a}$$

such that it says for every $x \in X$ whether this element x belongs to the set A or not:

$$f_A(x) = 1 \Leftrightarrow x \in A \quad \text{and} \quad f_A(x) = 0 \Leftrightarrow x \notin A. \tag{5.326b}$$

The concept of fuzzy sets is based on the idea of considering the membership of an element of the set as a statement, the truth value of which is characterized by a value from the interval $[0,1]$. For mathematical modeling of a fuzzy set A a function is necessary whose range is the interval $[0,1]$ instead of $\{0,1\}$, i.e.:

$$\mu_A\colon X \to [0,1]. \tag{5.327}$$

In other words: To every element $x \in X$ is to assign a number $\mu_A(x)$ from the interval $[0,1]$, which represents the grade of membership of x in A. The mapping μ_A is called the *membership function*. The value of the function $\mu_A(x)$ at the point x is called the *grade of membership*. The fuzzy sets A, B, C, etc. over X are also called fuzzy subsets of X. The set of all fuzzy sets over X is denoted by $F(X)$.

2. Properties of Fuzzy Sets and Further Definitions

The properties below follow directly from the definition:

(E1) Crisp sets can be interpreted as fuzzy sets with grade of membership 0 and 1.

(E2) The set of the arguments x, whose grade of membership is greater than zero, i.e., $\mu_A(x) > 0$, is called the *support* of the fuzzy set A:

$$\operatorname{supp}(A) = \{x \in X \mid \mu_A(x) > 0\}\,. \tag{5.328}$$

The set $ker(A) = \{x \in X : \mu_A(x) = 1\}$ is called the *kernel* or *core* of A.

(E3) Two fuzzy sets A and B over the universe X are equal if the values of their membership functions are equal:

$$A = B, \text{ if } \mu_A(x) = \mu_B(x) \text{ holds for every } x \in X. \tag{5.329}$$

(E4) Discrete representation or ordered pair representation: If the universe X is finite, i.e., $X = \{x_1, x_2, \ldots, x_n\}$ it is reasonable to define the membership function of the fuzzy set with a table of values. The tabular representation of the fuzzy set A is seen in **Table 5.7**.

Table 5.7 Tabular representation of a fuzzy set

x_1	x_2	$\ldots$	x_n
$\mu_A(x_1)$	$\mu_A(x_2)$	$\ldots$	$\mu_A(x_n)$

Also it is possible to write

$$A := \mu_A(x_1)/x_1 + \cdots + \mu_A(x_n)/x_n = \sum_{i=1}^{n} \mu_A(x_i)/x_i. \tag{5.330}$$

In (5.330) the fraction bars and addition signs have only symbolic meaning.

(E5) Ultra-fuzzy set: A fuzzy set, whose membership function itself is a fuzzy set, is called, after Zadeh, an *ultra-fuzzy set*.

3. Fuzzy Linguistics

Assigning linguistic values, e.g., “small”, “medium” or “big”, to a quantity then it is called a *linguistic quantity* or *linguistic variable.* Every linguistic value can be described by a fuzzy set, for example, by the graph of a membership function (5.9.1.2) with a given support (5.328). The number of fuzzy sets (in the case of “small”, “medium”, “big” they are three) depends on the problem.
In 5.9.1.2 the linguistic variable is denoted by x. For example, x can have linguistic values for temperature, pressure, volume, frequency, velocity, brightness, age, wearing, etc., and also medical, electrical, chemical, ecological, etc. variables.

■ By the membership function $\mu_A(x)$ of a linguistic variable, the membership degree of a fixed (crisp) value can be determined in the fuzzy set represented by $\mu_A(x)$. Namely, the modeling of a “high” quantity, e.g., the temperature, as a linguistic variable given by a trapezoidal membership function **(Fig. 5.65)** means that the given temperature α belongs to the fuzzy set “high temperature” with the degree of membership β (also degree of compatibility or degree of truth).

5.9.1.2 Membership Functions on the Real Line

The membership functions can be modeled by functions with values between 0 and 1. They represent the different grade of membership for the points of the universe being in the given set.

1. Trapezoidal Membership Functions

Trapezoidal membership functions are widespread. Piecewise (continuously differentiable) membership functions and their special cases, e.g., the triangle shape membership functions described in the following examples, are very often used. Connecting fuzzy quantities gives smoother output functions if the fuzzy quantities were represented by continuous or piecewise continuous membership functions.

■ **A:** Trapezoidal function **(Fig. 5.65)** corresponding to (5.331).

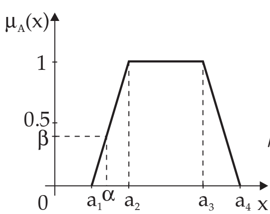


Figure 5.65

$$\mu_A(x) = \begin{cases} 0 & x \le a_1\,, \\ \dfrac{x-a_1}{a_2-a_1} & a_1 < x < a_2\,, \\ 1 & a_2 \le x \le a_3\,, \\ \dfrac{a_4-x}{a_4-a_3} & a_3 < x < a_4\,, \\ 0 & x \ge a_4. \end{cases} \qquad (5.331)$$

The graph of this function turns into a triangle function if $a_2 = a_3 = a$ and $a_1 < a < a_4$. Choosing different values for $a_1, \ldots, a_4$ gives symmetrical or asymmetrical trapezoidal functions, a symmetrical triangle function ($a_2 = a_3 = a$ and $|a - a_1| = |a_4 - a|$) or asymmetrical triangle function ($a_2 = a_3 = a$ and $|a - a_1| \neq |a_4 - a|$).

■ **B:** Membership function bounded to the left and to the right (**Fig. 5.66)** corresponding to (5.332):

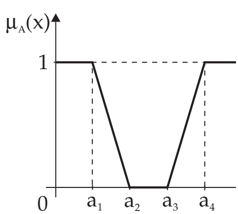


Figure 5.66

$$\mu_A(x) = \begin{cases} 1 & x \le a_1, \\ \dfrac{a_2-x}{a_2-a_1} & a_1 < x < a_2, \\ 0 & a_2 \le x \le a_3, \\ \dfrac{x-a_3}{a_4-a_3} & a_3 < x < a_4, \\ 1 & a_4 \le x. \end{cases} \qquad (5.332)$$

■ **C:** Generalized trapezoidal function **(Fig. 5.67)** corresponding to (5.333).

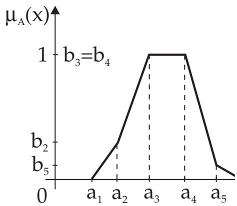


Figure 5.67

$$\mu_A(x) = \begin{cases} 0 & x \le a_1, \\ \dfrac{b_2(x-a_1)}{a_2-a_1} & a_1 < x < a_2, \\ \dfrac{(b_3-b_2)(x-a_2)}{a_3-a_2} + b_2 & a_2 \le x \le a_3, \\ b_3 = b_4 = 1 & a_3 < x < a_4, \\ \dfrac{(b_4-b_5)(a_4-x)}{a_5-a_4} + b_5 & a_4 < x \le a_5, \\ \dfrac{b_5(a_6-x)}{a_6-a_5} & a_5 < x < a_6, \\ 0 & a_6 \le x. \end{cases} \quad (5.333)$$

2. Bell-Shaped Membership Functions

■ **A:** A class of bell-shaped, differentiable membership functions is given by the function $f(x)$ from (5.334) by choosing an appropriate $p(x)$:

$$f(x) = \begin{cases} 0 & x \le a, \\ e^{-1/p(x)} & a < x < b, \\ 0 & x \ge b. \end{cases} \quad (5.334)$$

For $p(x) = k(x-a)(b-x)$ and, e.g., $k = 10$ or $k = 1$ or $k = 0.1$, there is a family of symmetrical curves of different width with the membership function $\mu_A(x) = f(x) \Big/ f\left(\dfrac{a+b}{2}\right)$, where $1 \Big/ f\left(\dfrac{a+b}{2}\right)$ is the normalizing factor **(Fig. 5.68)**. The exterior curve follows with the value $k = 10$ and the interior one with $k = 0.1$.

Asymmetrical membership functions in $[0, 1]$ follow e.g. for $p(x) = x(1-x)(2-x)$ or for $p(x) = x(1-x)(x+1)$ **(Fig. 5.69)**, using appropriate normalizing factors. The factor $(2-x)$ in the first polynomial results in the shifting of the maximum to the left and it yields an asymmetrical curve shape. Similarly, the factor $(x+1)$ in the second polynomial results in a shifting to the right and in an asymmetric form.

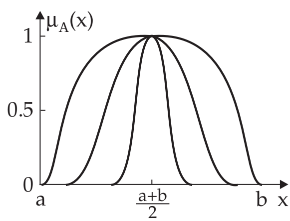


Figure 5.68

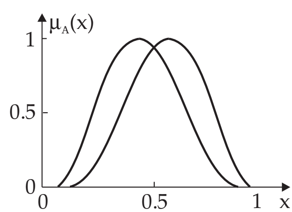


Figure 5.69

■ **B:** A more flexible class of membership functions can be got by the formula

$$F_t(x) = \frac{\displaystyle\int_a^x f\,(t(u))\; du}{\displaystyle\int_a^b f\,(t(u))\; du}, \quad (5.335)$$

where f is defined by (5.334) with $p(x) = (x-a)(b-x)$ and t is a transformation on $[a,b]$. If t is a smooth transformation on $[a,b]$, i.e., if t is differentiable infinitely many times in the interval $[a,b]$, then F_t is also smooth, since f is smooth. Requiring t to be either increasing or decreasing and to be smooth, then the transformation t allows to change the shape of the curve of the membership function. In practice, polynomials are especially suitable for transformations. The simplest polynomial is the identity $t(x) = x$ on the interval $[a,b] = [0,1]$.

The next simplest polynomial with the given properties is $t(x) = -\frac{2}{3}cx^3 + cx^2 + \left(1 - \frac{c}{3}\right)x$ with a constant $c \in [-6,3]$. The choice $c = -6$ results in the polynomial of maximum curvature, its equation is $q(x) = 4x^3 - 6x^2 + 3x$. Choosing for q_0 the identity function, i.e., $q_0(x) = x$, then can be got recursively further polynomials q by the formula $q_i = q \circ q_{i-1}$ for $i \in \mathbb{N}$. Substituting the corresponding polynomial transformations $q_0, q_1, \ldots$ into (5.335) for t, gives a sequence of smooth functions F_{q_0}, F_{q_1} and F_{q_2} **(Fig. 5.70)**, which can be considered as membership functions $\mu_A(x)$, where F_{q_n} converges to a line. The trapezoidal membership function can be approximated by differentiable functions using the function F_{q_2}, its reflection and a horizontal line **(Fig. 5.71)**.

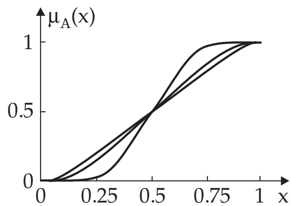


Figure 5.70

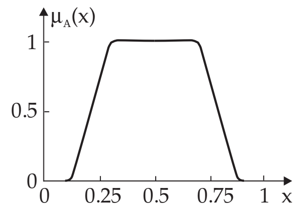


Figure 5.71

Summary: Imprecise and non-crisp information can be described by fuzzy sets and represented by membership functions $\mu(x)$.

5.9.1.3 Fuzzy Sets

1. Empty and Universal Fuzzy Sets

a) Empty fuzzy set: A set A over X is called *empty* if $\mu_A(x) = 0 \; \forall x \in X$ holds.

b) Universal fuzzy set: A set is called *universal* if $\mu_A(x) = 1 \; \forall x \in X$ holds.

2. Fuzzy Subset

If $\mu_B(x) \leq \mu_A(x) \; \forall \, x \in X$, then B is called a *fuzzy subset* of A (one writes: $B \subseteq A$).

3. Tolerance Interval and Spread of a Fuzzy Set on the Real Line

If A is a fuzzy set on the real line, then the interval

$$[a,b] = \{x \in X | \mu_A(x) = 1\} \quad (a, b \text{ const}, \; a < b) \tag{5.336}$$

is called the *tolerance interval* of the fuzzy set A, and the interval $[c,d] = \text{cl}(\text{supp}A)$ (c, d const,$c < d$) is called the *spread* of A, where cl denotes the closure of the set. (The tolerance interval is sometimes also called the *peak* of set A.) The tolerance interval and the kernel coincide only if the kernel contains more then one point.

■ **A:** In **Fig. 5.65** $[a_2, a_3]$ is the tolerance interval, and $[a_1, a_4]$ is the spread.

■ **B:** $a_2 = a_3 = a$ **(Fig. 5.65)**, gives a triangle-shaped membership function μ. In that case the triangular fuzzy set has no tolerance, but its kernel is the set $\{a\}$. If additionally $a_1 = a = a_4$ holds, too, then a crisp value follows; it is called a *singleton*. A singleton A has no tolerance, but $\ker(A) = \text{supp}(A) = \{a\}$.

4. Conversion of Fuzzy Sets on a Continuous and Discrete Universe

Let the universe be continuous, and let a fuzzy set be given on it by its membership function. Discretizing the universe, every discrete point together with its membership value determines a fuzzy singleton.

Conversely, a fuzzy set given on a discrete universe can be converted into a fuzzy set on the continuous universe by interpolating the membership value between the discrete points of the universe.

5. Normal and Subnormal Fuzzy Sets

If A is a fuzzy subset of X, then its *height* is defined by

$$H(A) := \max\{\mu_A(x)|x \in X\}. \tag{5.337}$$

A is called a *normal fuzzy set* if $H(A) = 1$, otherwise it is *subnormal*.
The notions and methods represented in this paragraph are limited to normal fuzzy sets, but it easy to extend them also to subnormal fuzzy sets.

6. Cut of a Fuzzy Set

The α *cut* $A^{>\alpha}$ or the *strong* α *cut* $A^{\geq\alpha}$ of a fuzzy set A are the subsets of X defined by

$$A^{>\alpha} = \{x \in X|\mu_A(x) > \alpha\}, \qquad A^{\geq\alpha} = \{x \in X|\mu_A(x) \geq \alpha\}, \quad \alpha \in (0,1]. \tag{5.338}$$

and $A^{\geq 0} = \mathrm{cl}\,(A^{>0})$. The α *cut* and *strong* α *cut* are also called α*-level set* and *strong* α*-level set*, respectively.

1. Properties

a) The α cuts of fuzzy sets are crisp sets.

b) The support $\mathrm{supp}(A)$ is a special α cut: $\mathrm{supp}(A) = A^{>0}$.

c) The crisp 1 cut $A^{\geq 1} = \{x \in X|\mu_A(x) = 1\}$ is called the *kernel* of A.

2. Representation Theorem

To every fuzzy subset A of X can be assigned uniquely the families of its α cuts $(A^{>\alpha})_{\alpha\in[0,1)}$ and its strong α cuts $\left(A^{\geq\alpha}\right)_{\alpha\in(0,1]}$. The α cuts and strong α cuts are monotone families of subsets from X, since:

$$\alpha < \beta \Rightarrow A^{>\alpha} \supseteq A^{>\beta} \quad \text{and} \quad A^{\geq\alpha} \supseteq A^{\geq\beta}. \tag{5.339a}$$

Conversely, if there exist the monotone families $(U_\alpha)_{\alpha\in[0,1)}$ or $(V_\alpha)_{\alpha\in(0,1]}$ of subsets from X, then there are uniquely defined fuzzy sets U and V such that $U^{>\alpha} = U_\alpha$ and $V^{\geq\alpha} = V_\alpha$ and moreover

$$\mu_U(x) = \sup\{\alpha \in [0,1))|x \in U_\alpha\}, \qquad \mu_V(x) = \sup\{\alpha \in (0,1]|x \in V_\alpha\}. \tag{5.339b}$$

7. Similarity of the Fuzzy Sets A and B

1. The fuzzy sets A, B with membership functions $\mu_A, \mu_B\colon X \to [0,1]$ are called fuzzy similar if for every $\alpha \in (0,1]$ there exist numbers α_i with $\alpha_i \in (0,1]; (i = 1,2)$ such that:

$$\mathrm{supp}(\alpha_1\mu_A)_\alpha \subseteq \mathrm{supp}(\mu_B)_\alpha, \qquad \mathrm{supp}(\alpha_2\mu_B)_\alpha \subseteq \mathrm{supp}(\mu_A)_\alpha. \tag{5.340}$$

$(\mu_C)_\alpha$ represents a fuzzy set with the membership function $(\mu_C)_\alpha = \begin{cases}\mu_C(x) & \text{if } \mu_C(x) > \alpha\\ 0 & \text{otherwise}\end{cases}$ and $(\beta\mu_C)$ represents a fuzzy set with the membership function $(\beta\mu_C) = \begin{cases}\beta \text{ if } \mu_C(x) > \beta\\ 0 \text{ otherwise.}\end{cases}$

2. Theorem: Two fuzzy sets A, B with membership functions $\mu_A, \mu_B\colon X \to [0,1]$ are fuzzy-similar if they have the same kernel:

$$\mathrm{supp}(\mu_A)_1 = \mathrm{supp}(\mu_B)_1, \tag{5.341a}$$

since the kernel is equal to the 1 cut, i.e.

$$\mathrm{supp}(\mu_A)_1 = \{x \in X|\mu_A(x) = 1\}. \tag{5.341b}$$

3. A, B with $\mu_A, \mu_B\colon X \to [0,1]$ are called strongly fuzzy-similar if they have the same support and the same kernel:

$$\text{supp}(\mu_A)_1 = \text{supp}(\mu_B)_1, \quad (5.342a) \qquad \text{supp}(\mu_A)_0 = \text{supp}(\mu_B)_0. \quad (5.342b)$$

5.9.2 Connections (Aggregations) of Fuzzy Sets

Fuzzy sets can be aggregated by operators. There are several different suggestions of how to generalize the usual set operations, such as union, intersection, and complement of fuzzy sets.

5.9.2.1 Concepts for Aggregations of Fuzzy Sets

1. Fuzzy Set Union, Fuzzy Set Intersection

The grade of membership of an arbitrary element $x \in X$ in the sets $A \cup B$ and $A \cap B$ should depend only on the grades of membership $\mu_A(x)$ and $\mu_B(x)$ of the element in the two fuzzy sets A and B. The union and intersection of fuzzy sets is defined with the help of two functions

$$s, t\colon [0,1] \times [0,1] \to [0,1], \quad (5.343)$$

and they are defined in the following way:

$$\mu_{A\cup B}(x) := s\left(\mu_A(x), \mu_B(x)\right), \quad (5.344) \qquad \mu_{A\cap B}(x) := t\left(\mu_A(x), \mu_B(x)\right). \quad (5.345)$$

The grades of membership $\mu_A(x)$ and $\mu_B(x)$ are mapped in a new grade of membership. The functions t and s are called the t norm and t conorm; this last one is also called the s norm.

Interpretation: The functions $\mu_{A\cup B}$ and $\mu_{A\cap B}$ represent the truth values of membership, which is resulted by the aggregation of the truth values of memberships $\mu_A(x)$ and $\mu_B(x)$.

2. Definition of the *t*Norm:

The t norm is a binary operation t in $[0,1]$:

$$t\colon [0,1] \times [0,1] \to [0,1]. \quad (5.346)$$

It is symmetric, associative, monotone increasing, it has 0 as the zero element and 1 as the neutral element. For $x, y, z, v, w \in [0,1]$ the following properties are valid:

(E1) Commutativity: $t(x,y) = t(y,x)$. (5.347a)

(E2) Associativity: $t(x, t(y,z)) = t(t(x,y), z)$. (5.347b)

(E3) Special Operations with Neutral and Zero Elements:

$t(x,1) = x$ and because of (E1): $t(1,x) = x$; $t(x,0) = t(0,x) = 0$. (5.347c)

(E4) Monotony: If $x \le v$ and $y \le w$, then $t(x,y) \le t(v,w)$ is valid. (5.347d)

3. Definition of the *s*Norm:

The s norm is a binary function in $[0,1]$:

$$s\colon [0,1] \times [0,1] \to [0,1]. \quad (5.348)$$

It has the following properties:

(E1) Commutativity: $s(x,y) = s(y,x)$. (5.349a)

(E2) Associativity: $s(x, s(y,z)) = s(s(x,y), z)$. (5.349b)

(E3) Special Operations with Zero and Neutral Elements:

$s(x,0) = s(0,x) = x$; $s(x,1) = s(1,x) = 1$. (5.349c)

(E4) Monotony: If $x \le v$ and $y \le w$, then $s(x,y) \le s(v,w)$ is valid. (5.349d)

With the help of these properties a class T of t norms and a class S of s norms can be introduced. Detailed investigations proved that the following relations hold:

$$\min\{x,y\} \ge t(x,y)\ \forall\, t \in T,\ \forall\, x,y \in [0,1] \quad \text{and} \quad (5.349e)$$

$$\max\{x,y\} \le s(x,y)\ \forall\, s \in S,\ \forall\, x,y \in [0,1]. \quad (5.349f)$$

5.9.2.2 Practical Aggregation Operations of Fuzzy Sets

1. Intersection of Two Fuzzy Sets

The *intersection* $A \cap B$ of two fuzzy sets A and B is defined by the minimum operation min(.,.) on their membership functions $\mu_A(x)$ and $\mu_B(x)$. Based on the previous requirements there is:

$$C := A \cap B \text{ and } \mu_C(x) := \min(\mu_A(x), \mu_B(x)) \quad \forall x \in X, \quad \text{where:} \tag{5.350a}$$

$$\min(a, b) := \begin{cases} a, & \text{if } a \le b, \\ b, & \text{if } a > b. \end{cases} \tag{5.350b}$$

The intersection operation corresponds to the AND operation of two membership functions **(Fig.5.72)**. The membership function $\mu_C(x)$ is defined as the minimum value of $\mu_A(x)$ and $\mu_B(x)$.

2. Union of Two Fuzzy Sets

The *union* $A \cup B$ of two fuzzy sets is defined by the maximum operation max(.,.) on their membership functions $\mu_A(x)$ and $\mu_B(x)$:

$$C := A \cup B \text{ and } \mu_C(x) := \max(\mu_A(x), \mu_B(x)) \quad \forall x \in X, \quad \text{where:} \tag{5.351a}$$

$$\max(a, b) := \begin{cases} a, & \text{if } a \ge b, \\ b, & \text{if } a < b. \end{cases} \tag{5.351b}$$

The union corresponds to the logical OR operation. **Fig.5.73** illustrates $\mu_C(x)$ as the maximum value of the membership functions $\mu_A(x)$ and $\mu_B(x)$.

■ The t norm $t(x, y) = \min\{x, y\}$ and the s norm $s(x, y) = \max\{x, y\}$ define the intersection and the union of two fuzzy sets, respectively (see **(Fig.5.74)** and **(Fig.5.75)**).

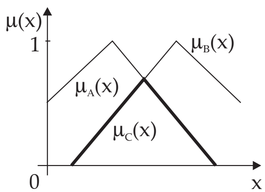


Figure 5.72

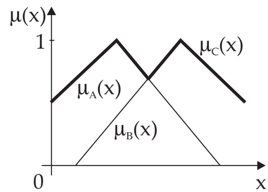


Figure 5.73

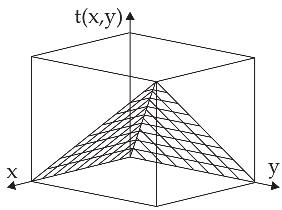


Figure 5.74

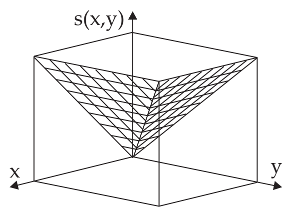


Figure 5.75

3. Further Aggregations

Further aggregations are the *bounded*, the *algebraic*, and the *drastic sum* and also the *bounded difference*, the *algebraic* and the *drastic product* (see **Table 5.8**).

The algebraic sum, e.g., is defined by

$$C := A + B \text{ and } \mu_C(x) := \mu_A(x) + \mu_B(x) - \mu_A(x) \cdot \mu_B(x) \quad \text{for every } x \in X. \tag{5.352a}$$

Similarly to the union (5.351a,b), this sum also belongs to the class of s norms. They are included in

Table 5.8 t and s norms, $p \in \mathbb{R}$

Author	**t norm**	**s norm**
Zadeh	intersection: $t(x,y) = \min\{x,y\}$	union: $s(x,y) = \max\{x,y\}$
Lukasiewicz	bounded difference $t_b(x,y) = \max\{0, x+y-1\}$	bounded sum $s_b(x,y) = \min\{1, x+y\}$
	algebraic product $t_a(x,y) = xy$	algebraic sum $s_a(x,y) = x+y-xy$
	drastic product $t_{dp}(x,y) = \begin{cases} \min\{x,y\}, \text{ whether } x=1 \text{ or } y=1 \\ 0 \text{ otherwise} \end{cases}$	drastic sum $s_{ds}(x,y) = \begin{cases} \max\{x,y\}, \text{ whether } x=0 \text{ or } y=0 \\ 1 \text{ otherwise} \end{cases}$
Hamacher ($p \geq 0$)	$t_h(x,y) = \dfrac{xy}{p+(1-p)(x+y-xy)}$	$s_h(x,y) = \dfrac{x+y-xy-(1-p)xy}{1-(1-p)xy}$
Einstein	$t_e(x,y) = \dfrac{xy}{1+(1-x)(1-y)}$	$s_e(x,y) = \dfrac{x+y}{1+xy}$
Frank ($p>0, p \neq 1$)	$t_f(x,y) = \log_p \left[1 + \dfrac{(p^x-1)(p^y-1)}{p-1}\right]$	$s_f(x,y) = 1 - \log_p \left[1 + \dfrac{(p^{1-x}-1)(p^{1-y}-1)}{p-1}\right]$
Yager ($p>0$)	$t_{ya}(x,y) = 1 - \min\left(1, ((1-x)^p + (1-y)^p)^{1/p}\right)$	$s_{ya}(x,y) = \min\left(1, (x^p+y^p)^{1/p}\right)$
Schweizer ($p>0$)	$t_s(x,y) = \max(0, x^{-p}+y^{-p}-1)^{-1/p}$	$s_s(x,y) = 1 - \max\left(0, (1-x)^{-p}+(1-y)^{-p}-1\right)^{-1/p}$
Dombi ($p>0$)	$t_{do}(x,y) = \left\{1 + \left[\left(\dfrac{1-x}{x}\right)^p + \left(\dfrac{1-y}{y}\right)^p\right]^{1/p}\right\}^{-1}$	$s_{do}(x,y) = 1 - \left\{1 + \left[\left(\dfrac{x}{1-x}\right)^p + \left(\dfrac{y}{1-y}\right)^p\right]^{1/p}\right\}^{-1}$
Weber ($p \geq -1$)	$t_w(x,y) = \max(0, (1+p) \cdot (x+y-1) - pxy)$	$s_w(x,y) = \min(1, x+y+pxy)$
Dubois ($0 \leq p \leq 1$)	$t_{du}(x,y) = \dfrac{xy}{\max(x,y,p)}$	$s_{du}(x,y) = \dfrac{x+y-xy-\min(x,y,(1-p))}{\max((1-x),(1-y),p)}$

Remark: For the values of the t and s norms listed in the table, the following ordering is valid:
$t_{dp} \leq t_b \leq t_e \leq t_a \leq t_h \leq t \leq s \leq s_h \leq s_a \leq s_e \leq s_b \leq s_{ds}$.

the right-hand column of **Table 5.8**. In **Table 5.9** is given a comparision of operations in Boolean logic and fuzzy logic.

Analogously to the notion of the extended sum as a union operation, the intersection can also be extended for example by the bounded, the algebraic, and the drastic product. So, e.g., the algebraic product is defined in the following way:

$$C := A \cdot B \text{ and } \mu_C(x) := \mu_A(x) \cdot \mu_B(x) \quad \text{for every } x \in X. \tag{5.352b}$$

It also belongs to the class of t norms, similarly to the intersection (5.350a,b), and it can be found in the middle column of **Table 5.8**.

5.9.2.3 Compensatory Operators

Sometimes operators are necessary lying between the t and the s norms; they are called compensatory operators. Examples for compensatory operators are the lambda and the gamma operator.

1. Lambda Operator

$$\mu_{A\lambda B}(x) = \lambda\,[\mu_A(x)\mu_B(x)] + (1-\lambda)\,[\mu_A(x) + \mu_B(x) - \mu_A(x)\mu_B(x)] \quad \text{with} \quad \lambda \in [0,1]. \quad (5.353)$$

Case $\lambda = 0$: Equation (5.353) results in a form known as the algebraic sum (**Table 5.8**, s norms); it belongs to the OR operators.

Case $\lambda = 1$: Equation (5.353) results in the form known as the algebraic product (**Table 5.8**, t norms); it belongs to the AND operators.

2. Gamma Operator

$$\mu_{A\gamma B}(x) = [\mu_A(x)\mu_B(x)]^{1-\gamma}\,[1-(1-\mu_A(x))\,(1-\mu_B(x))]^{\gamma} \quad \text{with } \gamma \in [0,1]. \quad (5.354)$$

Case γ**= 1:** Equation (5.354) results in the representation of the algebraic sum.

Case γ**= 0:** Equation (5.354) results in the representation of the algebraic product.

The application of the gamma operator on fuzzy sets of any numbers is given by

$$\mu(x) = \left[\prod_{i=1}^{n} \mu_i(x)\right]^{1-\gamma} \left[1 - \prod_{i=1}^{n}(1-\mu_i(x))\right]^{\gamma}, \quad (5.355)$$

and with weights δ_i:

$$\mu(x) = \left[\prod_{i=1}^{n} \mu_i(x)^{\delta_i}\right]^{1-\gamma} \left[1 - \prod_{i=1}^{n}(1-\mu_i(x))^{\delta_i}\right]^{\gamma} \quad \text{with } x \in X, \quad \sum_{i=1}^{n}\delta_i = 1, \quad \gamma \in [0,1]. \quad (5.356)$$

5.9.2.4 Extension Principle

In the previous paragraph there are discussed the possibilities of generalizing the basic set operations for fuzzy sets. Now, the notion of mapping is extended on fuzzy domains. The basis of the concept is the *acceptance grade* of vague statements. The classical mapping Φ: $X^n \to Y$ assigns a crisp function value $\Phi(x_1,\ldots,x_n) \in Y$ to the point $(x_1,\ldots,x_n) \in X^n$. This mapping can be extended for fuzzy variables as follows: The fuzzy mapping is $\hat{\Phi}$: $F(X)^n \to F(Y)$, which assigns a fuzzy function value $\hat{\Phi}(\mu_1,\ldots,\mu_n)$ to the fuzzy vector variables $(x_1,\ldots,x_n)$ given by the membership functions $(\mu_1,\ldots,\mu_n) \in F(X)^n$.

5.9.2.5 Fuzzy Complement

A function c: $[0,1] \to [0,1]$ is called a *complement function* if the following properties are fulfilled for $\forall\, x, y \in [0,1]$:

(EK1) Boundary Conditions: $c(0) = 1$ and $c(1) = 0$. (5.357a)

(EK2) Monotony: $x < y \Rightarrow c(x) \geq c(y)$. (5.357b)

(EK3) Involutivity: $c(c(x)) = x$. (5.357c)

(EK4) Continuity: $c(x)$ should be continuous for every $x \in [0,1]$. (5.357d)

■ **A:** The most often used complement function is (continuous and involutive):

$$c(x) := 1 - x. \quad (5.358)$$

■ **B:** Other continuous and involutive complements are the *Sugeno complement* $c_\lambda(x) := (1-x)(1+\lambda x)^{-1}$ with $\lambda \in (-1,\infty)$ and the *Yager complement* $c_p(x) := (1-x^p)^{1/p}$ with $p \in (0,\infty)$.

Table 5.9 Comparison of operations in Boolean logic and in fuzzy logic

Operator	**Boolean logic**	**Fuzzy logic** $(\boldsymbol{\mu_A, \mu_B \in [0,1]})$
AND	$C = A \wedge B$	$\mu_{A\cap B} = \min(\mu_A, \mu_B)$
OR	$C = A \vee B$	$\mu_{A\cup B} = \max(\mu_A, \mu_B)$
NOT	$C = \neg A$	$\mu_A^C = 1 - \mu_A$ (μ_A^C as complement of μ_A)

5.9.3 Fuzzy-Valued Relations

5.9.3.1 Fuzzy Relations

1. Modeling Fuzzy-Valued Relations

Uncertain or fuzzy-valued relations, as e.g. "approximately equal", "practically larger than", or "practically smaller than", etc., have an important role in practical applications. A relation between numbers is interpreted as a subsets of $\mathbb{R}^2$. So, the equality "=" is defined as the set

$$\mathcal{A} = \left\{(x,y) \in \mathbb{R}^2 | x = y\right\}, \tag{5.359}$$

i.e., by a straight line $y = x$ in $\mathbb{R}^2$.
Modeling the relation "approximately equal" denoted by R_1, can be used a fuzzy subset on $\mathbb{R}^2$, the kernel of which is $\mathcal{A}$. Furthermore it is to require that the membership function should decrease and tend to zero getting far from the line $\mathcal{A}$. A linear decreasing membership function can be modeled by

$$\mu_{R_1}(x,y) = \max\{0, 1 - a|x-y|\} \quad \text{with} \quad a \in \mathbb{R},\ a > 0. \tag{5.360}$$

For modeling the relation R_2 "practically larger than", it is useful to start with the crisp relation "$\geq$". The corresponding set of values is given by

$$\left\{(x,y) \in \mathbb{R}^2 | x \leq y\right\}. \tag{5.361}$$

It describes the crisp domain above the line $x = y$.
The modifier "practically" means that a thin zone under the half-space in (5.361) is still acceptable with some grade. So, the model of R_2 is

$$\mu_{R_2}(x,y) = \left\{ \begin{array}{ll} \max\{0, 1 - a|x-y|\} & \text{for } y < x \\ 1 & \text{for } y \geq x \end{array} \right\} \quad \text{with} \quad a \in \mathbb{R},\ a > 0. \tag{5.362}$$

If the value of one of the variables is fixed, e.g., $y = y_0$, then R_2 can be interpreted as a region with uncertain boundaries for the other variable.
Handling the uncertain boundaries by fuzzy relations has practical importance in fuzzy optimization, qualitative data analysis and pattern classification.

The foregoing discussion shows that the concept of fuzzy relations, i.e., fuzzy relations between several objects, can be described by fuzzy sets. In the following section the basic properties of binary relations are discussed over a universe which consists of ordered pairs.

2. Cartesian Product

Let X and Y be two universes. Their "cross product" $X \times Y$, or *Cartesian product*, is a universe G:

$$G = X \times Y = \{(x,y) | x \in X \wedge y \in Y\}. \tag{5.363}$$

Then, a fuzzy set on G is a fuzzy relation, analogously to classical set theory, if it consists of the valued pair of universes X and Y. A fuzzy relation R in G is a fuzzy subset $R \in F(G)$, where $F(G)$ denotes the set of all the fuzzy sets over $X \times Y$. R can be given by a membership function $\mu_R(x,y)$ which assigns a membership degree $\mu_R(x,y)$ from $[0,1]$ to every element of $(x,y) \in G$.

3. Properties of Fuzzy-Valued Relations

(E1) Since the fuzzy relations are special fuzzy sets, all propositions stated for fuzzy sets will also be valid for fuzzy relations.

(E2) All aggregations defined for fuzzy sets can be defined also for fuzzy relations; they yield a fuzzy

relation again.

(E3) The notion of α cut defined above can be transmitted without difficulties to fuzzy relations.

(E4) The 0 cut (the closure of the support) of a fuzzy relation $R \in F(G)$ is a usual relation on G.

(E5) Denoting the membership value by $\mu_R(x, y)$, i.e., the degree by which the relation R between the pair (x, y) holds. The value $\mu_R(x, y) = 1$ means that R holds perfectly for the pair (x, y), and the value $\mu_R(x, y) = 0$ means that R does not at all hold for the pair (x, y).

(E6) Let $R \in F(G)$ be a fuzzy relation. Then the fuzzy relation $S := R^{-1}$, the inverse of R, is defined by

$$\mu_S(x, y) = \mu_R(y, x) \quad \text{for every } (x, y) \in G. \tag{5.364}$$

■ The inverse relation R_2^{-1} means "practically smaller than" (see 5.9.3.1, **1.**, p. 422); the union $R_1 \cup R_2^{-1}$ can be determined as "practically smaller or approximately equal".

4. *n*-Fold Cartesian Product

Let n be the number of universal sets. Their *cross product* is an n-fold *Cartesian product.* A fuzzy set on an n-fold Cartesian product represents an n-fold fuzzy relation.

Consequences: The fuzzy sets, considered until now, are unary fuzzy relations, i.e., in the sense of the analysis they are curves above a universal set. A binary fuzzy relation can be considered as a surface over the universal set G. A binary fuzzy relation on a finite discrete support can be represented by a *fuzzy relation matrix.*

■ Colour-ripe grade relation: The well-known correspondence between the colour x and the ripe grade y of a friut is modeled in the form of a binary relation matrix with elements $\{0, 1\}$. The possible colours are $X = \{\text{green, yellow, red}\}$ and the ripe grades are $Y = \{\text{unripe, half-ripe, ripe}\}$. The relation matrix (5.365) belongs to the table:

	unripe	half-ripe	ripe
green	1	0	0
yellow	0	1	0
red	0	0	1

$$R = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{5.365}$$

Interpretation of this relation matrix: IF a fruit is green, THEN it is unripe. IF a fruit is yellow, THEN it is half-ripe. IF a fruit is red, THEN it is ripe. Green is uniquely assigned to unripe, yellow to half-ripe and red to ripe. If beyond it should be formalized that a green fruit can be considered half-ripe in a certain percentage, then the following table with discrete membership values can be arranged:

μ_R (green, unripe) $= 1.0$, μ_R (green, half-ripe) $= 0.5$,
μ_R (green, ripe) $= 0.0$, μ_R (yellow, unripe) $= 0.25$,
μ_R (yellow, half-ripe) $= 1.0$, μ_R (yellow, ripe) $= 0.25$,
μ_R (red, unripe) $= 0.0$, μ_R (red, half-ripe) $= 0.5$,
μ_R (red, ripe) $= 1.0$.

The relation matrix with $\mu_R \in [0, 1]$ is:

$$R = \begin{pmatrix} 1.0 & 0.5 & 0.0 \\ 0.25 & 1.0 & 0.25 \\ 0.0 & 0.5 & 1.0 \end{pmatrix}. \tag{5.366}$$

5. Rules of Calculations

The AND-type aggregation of fuzzy sets, e.g. $\mu_1 : X \to [0, 1]$ and $\mu_2 : Y \to [0, 1]$ given on different universes is formulated by the min operation as follows:

$$\mu_R(x, y) = \min(\mu_1(x), \mu_2(y)) \text{ or } (\mu_1 \times \mu_2)(x, y) = \min(\mu_1(x), \mu_2(y)) \quad \text{with} \tag{5.367a}$$

$$\mu_1 \times \mu_2 \colon G \to [0, 1], \text{ where } \quad G = X \times Y. \tag{5.367b}$$

The result of this aggregation is a fuzzy relation R on the cross product set (Cartesian product universe of fuzzy sets) G with $(x, y) \in G$. If X and Y are discrete finite sets and so $\mu_1(x), \mu_2(y)$ can be represented as vectors, then holds:

$$\mu_1 \times \mu_2 = \mu_1 \circ \mu_2^{\mathrm{T}} \quad \text{and} \quad \mu_{R^{-1}}(x, y) := \mu_R(y, x) \quad \forall\, (x, y) \in G. \tag{5.368}$$

The *aggregation operator* $\circ$ does not denote here the usual matrix product. The product is calculated here by the componentwise min operation and addition by the componentwise max operation.

The validity grade of an inverse relation R^{-1} for the pair (x, y) is always equal to the validity grade of R for the pair (y, x).

If the fuzzy relations are given on the same Cartesian product universe, then the rules of their aggregations can be given as follows: Let $R_1, R_2\colon X \times Y \to [0, 1]$ be binary fuzzy relations. The evaluation rule of their AND-type aggregation uses the min operator, namely for $\forall(x, y) \in G$:

$$\mu_{R_1 \cap R_2}(x, y) = \min(\mu_{R_1}(x, y), \mu_{R_2}(x, y)). \quad (5.369)$$

A corresponding evaluation rule for the OR-type aggregation is given by the max operation:

$$\mu_{R_1 \cup R_2}(x, y) = \max(\mu_{R_1}(x, y), \mu_{R_2}(x, y)). \quad (5.370)$$

5.9.3.2 Fuzzy Product Relation $R \circ S$

1. Composition or Product Relation

Suppose $R \in F(X \times Y)$ and $S \in F(Y \times Z)$ are two relations, and it is additionally assumed that $R, S \in F(G)$ with $G \subseteq X \times Z$. Then the *composition* or the *fuzzy product relation* $R \circ S$ is:

$$\mu_{R \circ S}(x, z) := \sup_{y \in Y}\{\min(\mu_R(x, y), \mu_S(y, z))\}\ \forall\,(x, z) \in X \times Z. \quad (5.371)$$

If a matrix representation is used for a finite universal set analogously to (5.366), then the composition $R \circ S$ is motivated as follows: Let $X = \{x_1, \dots, x_n\}, Y = \{y_1, \dots, y_m\}, Z = \{z_1, \dots, z_l\}$ and $R \in F(X \times Y),\ S \in F(Y \times Z)$ and let the matrix representations R, S be in the form $R = (r_{ij})$ and $S = (s_{jk})$ for $i = 1, \dots, n; j = 1, \dots, m; k = 1, \dots, l$, where

$$r_{ij} = \mu_R(x_i, y_j) \quad \text{and} \quad s_{jk} = \mu_S(y_j, z_k). \quad (5.372)$$

If the composition $T = R \circ S$ has the matrix representation t_{ik}, then

$$t_{ik} = \sup_j \min\{r_{ij}, s_{jk}\}. \quad (5.373)$$

The final result is not a usual matrix product, since instead of the summation operation there is the least upper bound (supremum) operation and instead of the product there is the minimum operator.

■ With the representations for r_{ij} and s_{jk} and with (5.371), the inverse relation $R^{-1}(r_{i,j})^{\mathrm{T}}$, can also be computed taking into consideration that R^{-1} can be represented by the transpose matrix, i.e., $R^{-1} = (r_{ij})^{\mathrm{T}}$.

Interpretation: Let R be a relation from X to Y and S be a relation from Y to Z. Then the following compositions are possible:

a) If the composition $R \circ S$ of R and S is defined as a max-min product, then the resulted fuzzy composition is called a max-min composition. The symbol sup stands for supremum and denotes the largest value, if no maximum exists.

b) If the product composition is defined as the usual matrix multiplication, then the max-prod composition is obtained.

c) For max-average composition, "multiplication" is replaced by the average.

2. Rules of Composition

The following rules are valid for the composition of fuzzy relations $R, S, T \in F(G)$:

(E1) Associative Law:

$$(R \circ S) \circ T = R \circ (S \circ T). \quad (5.374)$$

(E2) Distributive Law for Composition with Respect to the Union:

$$R \circ (S \cup T) = (R \circ S) \cup (R \circ T). \quad (5.375)$$

(E3) Distributive Law in a Weaker Form for Composition with Respect to Intersection:

$$R \circ (S \cap T) \subseteq (R \circ S) \cap (R \circ T). \quad (5.376)$$

(E4) Inverse Operations:

$$(R \circ S)^{-1} = S^{-1} \circ R^{-1}, \quad (R \cup S)^{-1} = R^{-1} \cup S^{-1} \quad \text{and} \quad (R \cap S)^{-1} = R^{-1} \cap S^{-1}. \quad (5.377)$$

(E5) Complement and Inverse:

$$\left(R^{-1}\right)^{-1} = R, \quad \left(R^{C}\right)^{-1} = \left(R^{-1}\right)^{C}. \tag{5.378}$$

(E6) Monotonic Properties:

$$R \subseteq S \Rightarrow R \circ T \subseteq S \circ T \quad \text{und } T \circ R \subseteq T \circ S. \tag{5.379}$$

■ **A:** Equation (5.371) for the product relation $R \circ S$ is defined by the min operation as we have done for intersection formation. In general, any t norm can be used instead of the min operation.

■ **B:** The α cuts with respect to the union, intersection, and complement are: $(A \cup B)^{>\alpha} = A^{>\alpha} \cup B^{>\alpha}$, $(A \cap B)^{>\alpha} = A^{>\alpha} \cap B^{>\alpha}$, $(A^C)^{>\alpha} = A^{\leq 1-\alpha} = \{x \in X | \mu_A(x) \leq 1 - \alpha\}$. Corresponding statements are valid for strong α cuts.

3. Fuzzy Logical Inferences

It is possible to make a fuzzy inference, e.g., with the IF THEN rule by the composition rule $\mu_2 = \mu_1 \circ R$. The detailed formulation for the conclusion μ_2 is given by

$$\mu_2(y) = \max_{x \in X}\Big(\min(\mu_1(x),\, \mu_R(x, y))\Big) \tag{5.380}$$

with $y \in Y,\ \mu_1\colon X \to [0, 1],\ \mu_2\colon Y \to [0, 1],\ R\colon G \to [0, 1]$ und $G = X \times Y$.

5.9.4 Fuzzy Inference (Approximate Reasoning)

Fuzzy inference is an application of fuzzy relations with the goal of getting fuzzy logical conclusions with respect to vague information (see 5.9.6.3, p. 428). Vague information means here fuzzy information but not uncertain information. Fuzzy inference, also called *implication*, contains one or more rules, a fact and a consequence. Fuzzy inference, which is called by Zadeh, approximate reasoning, cannot be described by classical logic.

1. Fuzzy Implication, IF THEN Rule

The fuzzy implication contains one IF THEN rule in the simplest case. The IF part is called the *premise* and it represents the condition. The THEN part is the *conclusion*. Evaluation happens by $\mu_2 = \mu_1 \circ R$ and (5.380).

Interpretation: μ_2 is the fuzzy inference image of μ_1 under the fuzzy relation R, i.e., a calculation prescription for the IF THEN rule or for a group of rules.

2. Generalized Fuzzy Inference Scheme

The rule IF A_1 AND A_2 AND $A_3 \ldots$ AND A_n THEN B with A_i: μ_i: $X_i \to [0, 1]$ $(i = 1, 2, \ldots, n)$ and the membership function of the conclusion B: μ: $Y \to [0, 1]$ is described by an $(n+1)$-valued relation

$$\text{R}\colon\ X_1 \times X_2 \times \cdots X_n \times Y \to [0, 1]. \tag{5.381a}$$

For the actual input with crisp values $x'_1, x'_2, \ldots, x'_n$ the rule (5.381a) defines the actual fuzzy output by

$$\mu_{B'}(y) = \mu_R(x'_1, x'_2, \ldots, x'_n, y) = \min(\mu_1(x'_1), \mu_2(x'_2), \ldots, \mu_n(x'_n), \mu_B(y)) \ \text{ where } \ y \in Y. \tag{5.381b}$$

Remark: The quantity $\min(\mu_1(x'_1),\ \mu_2(x'_2),\ \ldots\ \mu_n(x'_n))$ is called the *degree of fulfillment*, and the quantities $\{\mu_1(x'_1),\ \mu_2(x'_2),\ \ldots\ , \mu_n(x'_n)\}$ represent the fuzzy-valued input quantities.

■ Forming the fuzzy relations for a connection between the quantities “medium” pressure and “high” temperature **(Fig. 5.76)**: $\tilde{\mu}_1(p, T) = \mu_1(p)\ \forall T \in X_2$ with $\mu_1 : X_1 \to [0, 1]$ is a cylindrical extension **(Fig. 5.76c)** of the fuzzy set medium pressure **(Fig. 5.76a)**. Analogously, $\tilde{\mu}_2(p, T) = \mu_2(T)\ \forall p \in X_1$ with $\mu_2 :\ X_2 \to [0, 1]$ is a cylindrical extension **(Fig. 5.76d)** of the fuzzy set high temperature **(Fig. 5.76b)**, where $\tilde{\mu}_1, \tilde{\mu}_2\colon G = X_1 \times X_2 \to [0, 1]$.

Fig. 5.77a shows the graphic result of the formation of fuzzy relations: In **Fig. 5.77b** the result of the composition medium pressure AND high temperature with the min operator $\mu_R(p, T) = \min(\mu_1(p), \mu_2(T))$ is represented, and **(Fig. 5.77b)** shows the result of the composition OR with the max operator $\mu_R(p, T) = \max(\mu_1(p), \mu_2(T))$.

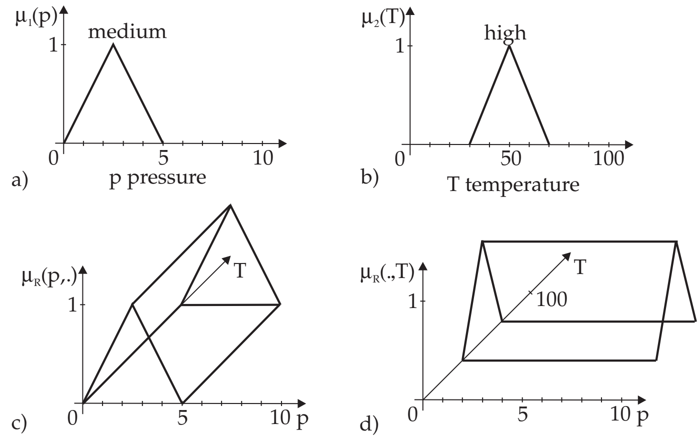


Figure 5.76

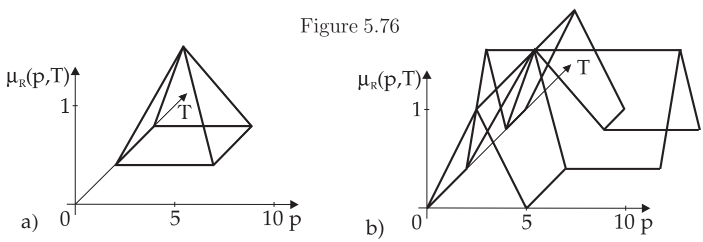


Figure 5.77

5.9.5 Defuzzification Methods

One has to get a crisp set from a fuzz-valued set in many cases. This process is called *defuzzification*. There are different methods available for this task.

1. Maximum-Criterion Method

An arbitrary value $\eta \in Y$ is selected from the domain where the fuzzy set $\mu^{\text{Output}}_{x_1,\ldots,x_n}$ has the maximal membership degree.

2. Mean-of-Maximum Method (MOM)

The output value is the mean value of the maximal membership values:

$$\sup\left(\mu^{\text{Output}}_{\mu_{x_1,\ldots,x_n}}\right) := \{y \in Y | \mu_{x_1,\ldots,x_n}(y) \geq \mu_{x_1,\ldots,x_n}(y^*)\ \forall\, y^* \in Y\}\ ; \quad (5.382)$$

i.e., the set Y is an interval, which should not be empty and it is characterized by (5.382), from which follows (5.383).

$$\eta_{\text{MOM}} = \frac{\displaystyle\int_{y \in \sup\left(\mu^{\text{Output}}_{x_1,\ldots,x_n}\right)} y\, dy}{\displaystyle\int_{y \in \sup\left(\mu^{\text{Output}}_{x_1,\ldots,x_n}\right)} dy} . \quad (5.383)$$

3. Center of Gravity Method (COG)

In the center of gravity method, one takes the abscissa value of the center of gravity of a surface with a fictitious homogeneous density of value 1.

$$\eta_{\text{COG}} = \frac{\displaystyle\int_{y_{\text{inf}}}^{y_{\text{sup}}} \mu(y) y\, dy}{\displaystyle\int_{y_{\text{inf}}}^{y_{\text{sup}}} \mu(y)\, dy} . \quad (5.384)$$

4. Parametrized Center of Gravity Method (PCOG)
The parametrized method works with the exponent $\gamma \in \mathbb{R}$. From (5.385) it follows for $\gamma = 1$, $\quad \eta_{\text{PCOG}} = \eta_{\text{COG}}$ and for $\gamma \to 0$, $\quad \eta_{\text{PCOG}} = \eta_{\text{MOM}}$.

$$\eta_{\text{PCOG}} = \frac{\int_{y_{\text{inf}}}^{y_{\text{sup}}} \mu(y)^{\gamma} y \, dy}{\int_{y_{\text{inf}}}^{y_{\text{sup}}} \mu(y)^{\gamma} \, dy}. \tag{5.385}$$

5. Generalized Center of Gravity Method (GCOG)
The exponent γ is considered as a function of y in the PCOG method. Then (5.386) follows obviously. The GCOG method is a generalization of the PCOG method, where $\mu(y)$ can be changed by the special weight γ depending itself on y.

$$\eta_{\text{GCOG}} = \frac{\int_{y_{\text{inf}}}^{y_{\text{sup}}} \mu(y)^{\gamma(y)} y \, dy}{\int_{y_{\text{inf}}}^{y_{\text{sup}}} \mu(y)^{\gamma(y)} \, dy}. \tag{5.386}$$

6. Center of Area (COA) Method
One calculates a line parallel to the ordinate axis so that the area under the membership function is the same on the left- and on the right-hand side of it.

$$\int_{y_{\text{inf}}}^{\eta} \mu(y) \, dy = \int_{\eta}^{y_{\text{sup}}} \mu(y) \, dy. \tag{5.387}$$

7. Parametrized Center of Area (PCOA) Method

$$\int_{y_{\text{inf}}}^{\eta_{\text{PB}}} \mu(y)^{\gamma} \, dy = \int_{\eta_{\text{PF}}}^{y_{\text{sup}}} \mu(y)^{\gamma} \, dy. \tag{5.388}$$

8. Method of the Largest Area (LA)
The significant subset is selected and one of the methods defined above, e.g., the method of center of gravity (COG) or center of area (COA) is used for this subset.

5.9.6 Knowledge-Based Fuzzy Systems

There are several application possibilities of multi-valued fuzzy logic, based on the unit interval, both in technical and non-technical life. The general concept consists in the fuzzification of quantities and characteristic numbers, in the aggregation them in an appropriate knowledge base with operators, and if necessary, in the defuzzification of the possibly fuzzy result set.

5.9.6.1 Method of Mamdani

The following steps are applied for a fuzzy control process:

1. Rule Base Suppose, for example, for the i-th rule

$$\mathrm{R}^i: \text{ If } e \text{ is } E^i \text{ AND } \dot{e} \text{ is } \Delta E^i \text{ THEN } u \text{ is } U^i. \tag{5.389}$$

Here e characterizes the error, $\dot{e}$ the change of the error and u the change of the (not fuzzy valued) output value. Every quantity is defined on its domain $E, \Delta E$ and U. Let the entire domain be $E \times \Delta E \times U$. The error and the change of the error will be fuzzified on this domain, i.e., they will be represented by fuzzy sets, where linguistic description is used.

2. Fuzzifying Algorithm In general, the error e and its change $\dot{e}$ are not fuzzy-valued, so they must be fuzzified by a linguistic description. The fuzzy values will be compared with the premisses of the IF THEN rule from the rule base. From this it follows, which rules are active and how large are their weights.

3. Aggregation Module The active rules with their different weights will be combined with an algebraic operation and applied to the defuzzification.

4. Decision Module In the defuzzification process a crisp value should be given for the control quantity. With a defuzzification operation, a non-fuzzy-valued quantity is determined from the set of possible values, i.e., a crisp quantity. This quantity expresses how the control parameters of the system should be set up to keep the deviation minimal.

Fuzzy control means that the steps from **1.** to **4.** are repeated until the goal, the smallest deviation e and its change $\dot{e}$, is reached.

5.9.6.2 Method of Sugeno

The Sugeno method is also used for planning of a fuzzy control process. It differs from the Mamdani concept in the rule base and in the defuzzification method. It has the following steps:

1. Rule Base: The rule base consists of rules of the following form:

$$R^i: \text{ IF } x_1 \text{ is } A_1^i \text{ AND } \ldots \text{ AND } x_k \text{ is } A_k^i, \text{THEN } u_i = p_0^i + p_1^i x_1 + p_2^i x_2 + \cdots + p_k^i x_k. \tag{5.390}$$

The notations mean:

A_j: fuzzy sets, which can be determined by membership functions;

x_j: crisp input values as, e.g., the error e and the change of the error $\dot{e}$, which tell us something about the dynamics of the system;

p_j^i: weights of x_j $(j = 1, 2, \ldots, k)$;

u_i: the output value belonging to the i-th rule $(i = 1, 2, \ldots, n)$.

2. Fuzzifying Algorithm: A $\mu_i \in [0, 1]$ is calculated for every rule R^i.

3. Decision Module: A non-fuzzy-valued quantity is calculated from the weighted mean of u_i, where the weights are μ_i from the fuzzification:

$$u = \sum_{i=1}^{n} \mu_i u_i \left(\sum_{i=1}^{n} \mu_i \right)^{-1}. \tag{5.391}$$

Here u is a crisp value.

The defuzzification of the Mamdani method does not work here. The problem is to get the weight parameters p_j^i available. These parameters can be determined by a mechanical learning method, e.g., by an artificial neuronetwork (ANN).

5.9.6.3 Cognitive Systems

To clarify the method, the following known example will be investigated with the Mamdami method: The regulation of a pendulum that is perpendicular to its moving base **(Fig. 5.78)**. The aim of the control process is to keep a pendulum in balance so that the pendulum rod should stand vertical, i.e., the angular displacement from the vertical direction and the angular velocity should be zero. It must be done by a force F acting at the lower end of the pendulum. This force is the control quantity. The model is based on the activity of a human "control expert" (cognitive problem). The expert formulates its knowledge in linguistic rules. Linguistic rules consist, in general, of a premise, i.e., a specification of the measured values, and a conclusion which gives the appropriate control value.

For every set of values $X_1, X_2, \ldots, X_n$ for the measured values and Y for the control quantity the appropriate linguistic terms are defined as "approximately zero", "small positive", etc. Here "approximately zero" with respect to the measured value ξ_1 can have a different meaning as for the measured value ξ_2.

■ **Inverse Pendulum on a Moving Base (Fig. 5.78)**

1. Modeling For the set X_1 (values of angle) and analogously for the input quantity X_2 (values of the angular velocity) the seven linguistic terms, negative large (nl), negative medium (nm), negative small (ns), zero (z), positive small (ps), positive medium (pm) and positive large (pl) are chosen.

For the mathematical modeling, a fuzzy set must be assigned by graphs to every one of these linguistic terms **(Fig. 5.77)**, as was shown for fuzzy inference (see 5.9.4, p. 425).

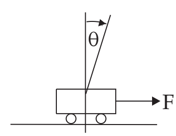


Figure 5.78

2. Choice of the Domain of Values

- Values of angles: $\Theta(-90° < \Theta < 90°)$: $X_1 := [-90°, 90°]$.
- Values of angular velocity: $\dot{\Theta}(-45°\,\mathrm{s}^{-1} \le \dot{\Theta} \le 45°\,\mathrm{s}^{-1})$: $X_2 := [-45°\,\mathrm{s}^{-1}, 45°\,\mathrm{s}^{-1}]$.
- Values of force F: $(-10\,\mathrm{N} \le F \le 10\,\mathrm{N})$: $Y := [-10\mathrm{N}, 10\,\mathrm{N}]$.

The partitioning of the input quantities X_1 and X_2 and the output quantity Y is represented graphically in **Fig. 5.79**. Usually, the initial values are actual measured values, e.g., $\Theta = 36°$, $\dot{\Theta} = -2.25°\,\mathrm{s}^{-1}$.

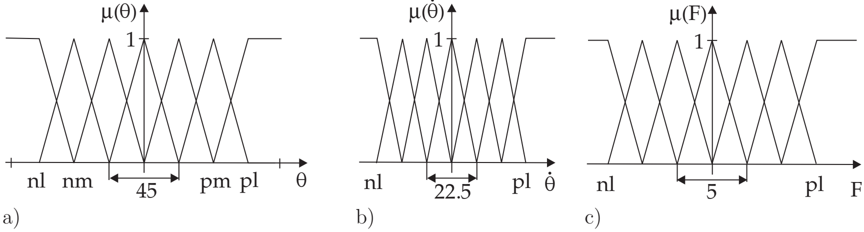


Figure 5.79

3. Choice of Rules Considering the following table, there are 49 possible rules (7×7) but there are only 19 important in practice, so the following two are to be discussed: **R1** and **R2**.

R1: If Θ is positive small (ps) and $\dot{\Theta}$ zero (z), then F is positive small (ps). For the *degree of fulfillment* (also called the *weight of the rules*) of the premise with $\alpha = \min\left\{\mu^{(1)}(\Theta); \mu^{(1)}(\dot{\Theta})\right\} = \min\{0.4; 0.8\} = 0.4$ one gets the output set (5.392) by an α cut, hence the output fuzzy set is positive small (ps) in the height $\alpha = 0.4$ **(Fig. 5.80c)**.

Table: Rule base with 19 practically meaningful rules

$\dot{\Theta}\backslash\Theta$	nl	nm	ns	z	ps	pm	pl
nl			ps	pl			
nm				pm			
ns	nm		ns	ps			
z	nl	nm	ns	z	ps	pm	pl
ps				ns	ps		pm
pm				nm			
pl				nl	ns		

$$\mu_{36;-2.25}^{\text{Output (R1)}}(y) = \begin{cases} \frac{2}{5}y & 0 \le y < 1, \\ 0.4 & 1 \le y \le 4, \\ 2 - \frac{2}{5}y & 4 < y \le 5, \\ 0 & \text{otherwise.} \end{cases} \tag{5.392}$$

R2: If Θ is positive medium (pm) and $\dot{\Theta}$ is zero (z), then F is positive medium (pm). For the performance score of the premise follows $\alpha = \min\left\{\mu^{(2)}(\Theta); \mu^{(2)}\dot{\Theta}\right\} = \min\{0.6; 0.8\} = 0.6$, the output set (5.393) analogously to rule **R1 (Fig. 5.80f)**.

$$\mu_{36;-2.25}^{\text{Output (R2)}}(y) = \begin{cases} \frac{2}{5}y - 1 & 2.5 \le y < 4, \\ 0.6 & 4 \le y \le 6, \\ 3 - \frac{2}{5}y & 6 < y \le 7.5, \\ 0 & \text{otherwise.} \end{cases} \tag{5.393}$$

4. Decision Logic The evaluation of rule R_1 with the min operation results in the fuzzy set in **Figs. 5.80a–c**. The corresponding evaluation for the rule R_2 is shown in **Figs. 5.80d–f**. The control quantity is calculated finally by a defuzzification method from the fuzzy proposition set **(Fig. 5.80g)**. The result is the fuzzy set **(Fig. 5.80g)** by using the max operation and taking into account the fuzzy sets **(Fig. 5.80c)** and **(Fig. 5.80f)**.

a) Evaluation of the fuzzy set obtained in this way, which is aggregated by operators (see max-min composition 5.9.3.2, **1.**, p. 424). The decision logic yields:

$$\mu_{x_1,\ldots,x_n}^{\text{Output}} : Y \to [0,1]\,; y \to \max\nolimits_{r\in\{1,\ldots,k\}} \left\{\min\left\{\mu_{i_{l,r}}^{(1)}(x_1), \ldots, \mu_{i_{l,r}}^{(n)}(x_n), \mu_{i_r}(y)\right\}\right\}. \tag{5.394}$$

b) After taking the maximum (5.395) is obtained for the function graph of the fuzzy set.

c) For the other 17 rules results a degree of fulfillment equal to zero for the premise, i.e., it results in fuzzy sets, which are zeros themselves.

5. Defuzzification The decision logic yields no crisp value for the control quantity, but a fuzzy set. That means, by this method, one gets a mapping, which assigns a fuzzy set $\mu^{\text{Output}}_{x_1,\ldots,x_n}$ of Y to every tuple $(x_1,\ldots,x_n) \in X_1 \times X_2 \times \cdots \times X_n$ of the measured values.

$$\mu^{\text{Output}}_{36;-2.25}(y) = \begin{cases} \frac{2}{5}y & \text{for} \quad 0 \le y < 1, \\ 0.4 & \text{for} \quad 1 \le y < 3.5, \\ \frac{2}{5}y - 1 & \text{for} \quad 3.5 \le y < 4, \\ 0.6 & \text{for} \quad 4 \le y < 6, \\ 3 - \frac{2}{5}y & \text{for} \quad 6 \le y \le 7.5, \\ 0 & \text{for} \quad \text{otherwise.} \end{cases} \qquad (5.395)$$

Defuzzification means that there is to determine a control quantity using defuzzification methods. The center of gravity method and the maximum criterion method result in the value for control quantity $F = 3.95$ or $F = 5.0$.

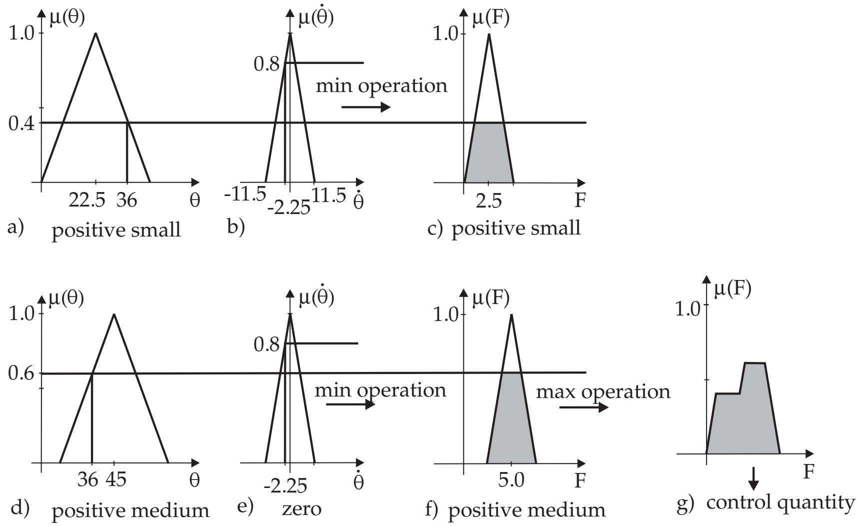


Figure 5.80

6. Remarks

1. The “knowledge-based” trajectories should lie in the rule base so that the endpoint is in the center of the smallest rule deviation.

2. By defuzzification an iteration process is introduced, which leads finally to the center of the partition space, i.e., which results in a zero control quantity.

3. Every non-linear domain of characteristics can be approximated with arbitrary accuracy by the choice of appropriate parameters on a compact domain.

5.9.6.4 Knowledge-Based Interpolation Systems

1. Interpolation Mechanism

Interpolation mechanisms can be built up with the help of fuzzy logic. Fuzzy systems are systems

to process fuzzy information. With them it is possible to approximate and interpolate functions. A simple fuzzy system, by which this property can be investigated , is the Sugeno controller. It has n input variables $\xi_1, \dots, \xi_n$ and defines the value of the output variable y by rules $\mathrm{R}_1, \dots, \mathrm{R}_n$ in the form

$$\mathrm{R}_i\colon \text{ IF } \xi_1 \text{ is } A_1^{(i)} \text{ and } \cdots \text{ and } \xi_n \text{ is } A_n^{(i)}, \text{THEN is } y = f_i(\xi_1, \dots, \xi_n) \quad (i = 1, 2, \dots, n). \tag{5.396}$$

The fuzzy sets $A_j^{(1)}, \dots, A_j^{(k)}$ always partition the input sets X_j. The conclusions $f_i(\xi_1, \dots, \xi_n)$ of the rules are singletons, which can depend on the input variables $\xi_1, \dots, \xi_n$.

By a simple choice of the conclusions the expensive defuzzification can be omitted and the output value y will be calculated as a weighted sum. To do this, the controller calculates a degree of fulfillment α_i for every rule R_i with a t norm from the membership grades of the single inputs and determines the output value

$$y = \frac{\sum_{i=1}^{N} \alpha_i f_i(\xi_1, \dots, \xi_n)}{\sum_{i=1}^{N} \alpha_i}. \tag{5.397}$$

2. Restriction to the One-Dimensional Case

For fuzzy systems with only one input $x = \xi_1$, fuzzy sets represented by triangular functions are often used which are cut at the height 0.5. Such fuzzy sets satisfy the following three conditions:

1. For every rule R_i there is an input x_i, for which only one rule is fulfilled. For this input x_i, the output is calculated by f_i. By this, the output of the fuzzy system is fixed at N nodes $x_1, \dots, x_N$. Actually, the fuzzy system interpolates the nodes $x_1, \dots, x_N$. The requirement that at the node x_i only one rule R_i holds, is sufficient for an exact interpolation, but it is not necessary. For two rules R_1 and R_2, as they will be considered below, this requirement means that $\alpha_1(x_2) = \alpha_2(x_1) = 0$ holds. To fulfill the first condition, $\alpha_1(x_2) = \alpha_2(x_1) = 0$ must hold. This is a sufficient condition for an exact interpolation of the nodes.

2. There are at most two rules fulfilled between two consecutive nodes. If x_1 and x_2 are two such nodes with rules R_1 and R_2, then for inputs $x \in [x_1, x_2]$ the output y is

$$y = \frac{\alpha_1(x) f_1(x) + \alpha_2(x) f_2(x)}{\alpha_1(x) + \alpha_2(x)} = f_1(x) + g(x)\,[f_2(x) - f_1(x)] \quad \text{with } g := \frac{\alpha_2(x)}{\alpha_1(x) + \alpha_2(x)}. \tag{5.398}$$

The actual shape of the interpolation curve between x_1 and x_2 is determined by the function g. The shape depends only on the satisfaction grades α_1 and α_2, which are the values of the membership functions $\mu_{A_i^{(1)}}$ and $\mu_{A_i^{(2)}}$ at the point x, i.e., $\alpha_1 = \mu_{A^{(1)}}(x)$ and $\alpha_2 = \mu_{A^{(2)}}(x)$ are valid, or in short form $\alpha_1 = \mu_1(x)$ and $\alpha_2 = \mu_2(x)$. The shape of the curve depends only on the relation μ_1/μ_2 of the membership functions.

3. The membership functions are positive, so the output y is a convex combination of the conclusions f_i. For the given and for the general case hold (5.399) and (5.400), respectively:

$$\min(f_1, f_2) \le y \le \max(f_1, f_2), \tag{5.399}$$

$$\min_{i\in\{1,2,\dots,N\}} f_i \le y \le \max_{i\in\{1,2,\dots,N\}} f_i. \tag{5.400}$$

For constant conclusions, the terms f_1 and f_2 cause only a translation and stretching of the shape of the curve g. If the conclusions are dependent on the input variables, then the shape of the curve is differently perturbed in different sections. Consequently, another output function can be found.

Applying linearly dependent conclusions and membership functions with constant sum for the input x, then the output is $y = c\sum\limits_{i=1}^{N} \alpha_i(x) f_i(x)$ with α_i depending on x and a constant c, so that the interpolation functions are polynomials of second degree. These polynomials can be used for the construction of an interpolation method with polynomials of second degree.

In general, choosing polynomials of n-th degree, an interpolation polynomial of $(n + 1)$-th degree is obtained as a conclusion. In this sense fuzzy systems are rule-based interpolation systems besides conventional interpolation methods interpolating locally by polynomials, e.g., with splines.

6 Differentiation

6.1 Differentiation of Functions of One Variable

6.1.1 Differential Quotient

1. Differential Quotient or Derivative of a Function

The *differential quotient* of a function $y = f(x)$ at x_0 is equal to $\lim\limits_{\Delta x \to 0} \dfrac{f(x_0 + \Delta x) - f(x_0)}{\Delta x}$ if this limit exists and is finite. The *derivative function* of a function $y = f(x)$ with respect to the variable x is another function of x denoted by the symbols y', $\dot{y}$, Dy, $\dfrac{dy}{dx}$, $f'(x)$, $Df(x)$, or $\dfrac{df(x)}{dx}$, and its value for every x is equal to the limit of the quotient of the increment of the function Δy and the corresponding increment Δx for $\Delta x \to 0$, if this limit exists:

$$f'(x) = \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x}. \tag{6.1}$$

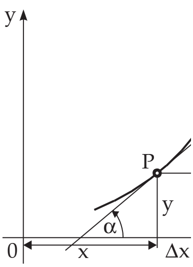


Figure 6.1

2. Geometric Representation of the Derivative

If $y = f(x)$ is represented as a curve in a Cartesian coordinate system as in **Fig. 6.1**, and if the x-axis and the y-axis have the same unit, then

$$f'(x) = \tan\alpha \tag{6.2}$$

is valid. The angle α between the x-axis and the tangent line of the curve at the considered point defines the *angular coefficient* or *slope of the tangent* (see 3.6.1.2, **2.**, p. 245). The angle is measured from the positive x-axis to the tangent in a counterclockwise direction, and it is called the *angle of slope* or *angle of inclination.*

3. Differentiability

From the definition of the derivative it obviously follows that $f(x)$ is differentiable with respect to x for the values of x where the differential quotient (6.1) has a finite value. The domain of the derivative function is a subset (proper or trivial) of the domain of the original function. If the function is continuous at x but the derivative does not exist, then perhaps there is no determined tangent line at that point, or the tangent line is perpendicular to the x-axis. In this last case the limit in (6.1) is infinity. For this case is used the notation $f'(x) = +\infty$ or $-\infty$.

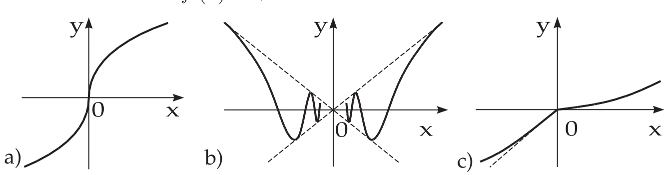


Figure 6.2

■ **A:** $f(x) = \sqrt[3]{x}$: $f'(x) = \dfrac{1}{3\sqrt[3]{x^2}}$, $f'(0) = \infty$. At the point 0 the limit (6.1) tends to infinity, so the derivative does not exist at the point 0 **(Fig. 6.2a)**.

■ **B:** $f(x) = x \sin\dfrac{1}{x}$ for $x \neq 0$. At the point $x = 0$ the function $f(x)$ is not defined, but it has zero limit, so one writes $f(0) = 0$. However the limit (6.1) does not exist at $x = 0$ **(Fig. 6.2b)**.

I.N. Bronshtein et al., *Handbook of Mathematics*,
DOI 10.1007/978-3-662-46221-8_6

4. Left-Hand and Right-Hand Differentiability

If the limit (6.1) does not exist for a value $x = a$, but the left-hand limit or the right-hand limit exists, this limit is called the left-hand derivative or right-hand derivative respectively. If both exist, the curve has two tangents here:

$$f'(a-0) = \tan\alpha_1, \qquad f'(a+0) = \tan\alpha_2. \tag{6.3}$$

Geometrically this means that the curve has a knee **(Fig. 6.2c, Fig. 6.3)**.

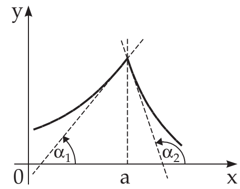


Figure 6.3

■ $f(x) = \dfrac{x}{1+e^{\frac{1}{x}}}$ for $x \neq 0$. For $x = 0$ the function is not defined, but it has zero limit at $x = 0$, so one writes $f(x) = 0$. At the point $x = 0$ there is no limit of type (6.1), but there is a left-hand and a right-hand limit $f'(-0) = 1$ and $f'(+0) = 0$, i.e., the curve has a knee here **(Fig. 6.2c)**.

6.1.2 Rules of Differentiation for Functions of One Variable

6.1.2.1 Derivatives of the Elementary Functions

The elementary functions have a derivative on all their domains except perhaps some points, as represented in **Fig. 6.2**.

A summary of the derivatives of elementary functions can be found in **Table 6.1**. Further derivatives of elementary functions can be found by reversing the results of the indefinite integrals in **Table 8.1**.

Remark: In practice, it is often useful to transform the function into a more convenient form to perform differentiation, e.g., to transform it into a sum where parentheses are removed (see 1.1.6.1, p. 11) or to separate the integral rational part of the expression (see 1.1.7, p. 14) or to take the logarithm of the expression (see 1.1.4.3, p. 9).

■ **A:** $y = \dfrac{2 - 3\sqrt{x} + 4\sqrt[3]{x} + x^2}{x} = \dfrac{2}{x} - 3x^{-\frac{1}{2}} + 4x^{-\frac{2}{3}} + x\,; \quad \dfrac{dy}{dx} = -2x^{-2} + \dfrac{3}{2}x^{-\frac{3}{2}} - \dfrac{8}{3}x^{-\frac{5}{3}} + 1\,.$

■ **B:** $y = \ln\sqrt{\dfrac{x^2+1}{x^2-1}} = \dfrac{1}{2}\ln(x^2+1) - \dfrac{1}{2}\ln(x^2-1)\,; \quad \dfrac{dy}{dx} = \dfrac{1}{2}\left(\dfrac{2x}{x^2+1}\right) - \dfrac{1}{2}\left(\dfrac{2x}{x^2-1}\right) = -\dfrac{2x}{x^4-1}\,.$

6.1.2.2 Basic Rules of Differentiation

Assume u, v, w, and y are functions of the independent variable x, and u', v', w', and y' are the derivatives with respect to x. The differential is denoted by du, dv, dw, and dy (see 6.2.1.3, p. 446). The basic rules of differentiation, which are explained separately, are summarized in **Table 6.2**, p. 439.

1. Derivative of a Constant Function

The derivative of a constant function c is the zero function:

$$c' = 0. \tag{6.4}$$

2. Derivative of a Scalar Multiple

A constant factor c can be factored out from the differential sign:

$$(c\,u)' = c\,u', \quad d(c\,u) = c\,du. \tag{6.5}$$

3. Derivative of a Sum

If the functions u, v, w, etc. are differentiable one by one, their sum and difference is also differentiable, and equal to the sum or difference of the derivatives:

$$(u + v - w)' = u' + v' - w', \tag{6.6a}$$

$$d(u + v - w) = du + dv - dw. \tag{6.6b}$$

It is possible that the summands are not differentiable separately, but their sum or difference is. Then the derivative must be calculated by definition formula (6.1).

Table 6.1 Derivatives of elementary functions in the intervals on which they are defined and the occurring numerators are not equal to zero

Function	**Derivative**	**Function**	**Derivative**
C (constant)	0	$\sec x$	$\dfrac{\sin x}{\cos^2 x}$
x	1	$\operatorname{cosec} x$	$\dfrac{-\cos x}{\sin^2 x}$
$x^n \quad (n \in \mathbb{R})$	nx^{n-1}	$\arcsin x \quad (\|x\| < 1)$	$\dfrac{1}{\sqrt{1-x^2}}$
$\dfrac{1}{x} \quad (x \neq 0)$	$-\dfrac{1}{x^2} \quad (x \neq 0)$	$\arccos x \quad (\|x\| < 1)$	$-\dfrac{1}{\sqrt{1-x^2}}$
$\dfrac{1}{x^n} \quad (x \neq 0)$	$-\dfrac{n}{x^{n+1}}$	$\arctan x$	$\dfrac{1}{1+x^2}$
$\sqrt{x} \quad (x > 0)$	$\dfrac{1}{2\sqrt{x}}$	$\operatorname{arccot} x$	$-\dfrac{1}{1+x^2}$
$\sqrt[n]{x} \quad (n \in \mathbb{R},\ n \neq 0,\ x > 0)$	$\dfrac{1}{n\sqrt[n]{x^{n-1}}}$	$\operatorname{arcsec} x \quad (x > 1)$	$\dfrac{1}{x\sqrt{x^2-1}}$
e^x	e^x	$\operatorname{arccosec} x \quad (x > 1)$	$-\dfrac{1}{x\sqrt{x^2-1}}$
$e^{bx} \quad (b \in \mathbb{R})$	be^{bx}	$\sinh x$	$\cosh x$
$a^x \quad (a > 0)$	$a^x \ln a$	$\cosh x$	$\sinh x$
$a^{bx} \quad (b \in \mathbb{R},\ a > 0)$	$ba^{bx} \ln a$	$\tanh x$	$\dfrac{1}{\cosh^2 x}$
$\ln x \quad (x > 0)$	$\dfrac{1}{x}$	$\coth x \quad (x \neq 0)$	$-\dfrac{1}{\sinh^2 x}$
$\log_a x \quad (a > 0,\ a \neq 1,\ x > 0)$	$\dfrac{1}{x} \log_a e = \dfrac{1}{x \ln a}$	$\operatorname{Arsinh} x$	$\dfrac{1}{\sqrt{1+x^2}}$
$\lg x \quad (x > 0)$	$\dfrac{1}{x} \lg e \approx \dfrac{0.4343}{x}$	$\operatorname{Arcosh} x \quad (x > 1)$	$\dfrac{1}{\sqrt{x^2-1}}$
$\sin x$	$\cos x$	$\operatorname{Artanh} x \quad (\|x\| < 1)$	$\dfrac{1}{1-x^2}$
$\cos x$	$-\sin x$	$\operatorname{Arcoth} x \quad (\|x\| > 1)$	$-\dfrac{1}{x^2-1}$
$\tan x \quad (x \neq (2k+1)\dfrac{\pi}{2},\ k \in \mathbb{Z})$	$\dfrac{1}{\cos^2 x} = \sec^2 x$	$[f(x)]^n \quad (n \in \mathbb{R})$	$n[f(x)]^{n-1} f'(x)$
$\cot x \quad (x \neq k\pi,\ k \in \mathbb{Z})$	$\dfrac{-1}{\sin^2 x} = -\operatorname{cosec}^2 x$	$\ln f(x) \quad (f(x) > 0)$	$\dfrac{f'(x)}{f(x)}$

4. Derivative of a Product

If two, three, or n functions are differentiable one by one, then their product is differentiable, and can be calculated as follows:

a) Derivative of the Product of Two Functions:

$$(u\,v)' = u'\,v + u\,v', \quad d(u\,v) = v\,du + u\,dv. \tag{6.7a}$$

It is possible that the terms are not differentiable separately, but their product is. Then the derivative must be calculated by definition formula (6.1).

b) Derivative of the Product of Three Functions:

$$(u\,v\,w)' = u'\,v\,w + u\,v'\,w + u\,v\,w', \quad d(u\,v\,w) = v\,w\,du + u\,w\,dv + u\,v\,dw. \tag{6.7b}$$

c) Derivative of the Product of n Functions:

$$(u_1\,u_2\,\cdots\,u_n)' = \sum_{i=1}^{n} u_1\,u_2\,\cdots\,u_i'\,\cdots\,u_n. \tag{6.7c}$$

■ **A:** $y = x^3 \cos x, \quad y' = 3x^2 \cos x - x^3 \sin x.$

■ **B:** $y = x^3 e^x \cos x, \quad y' = 3x^2 e^x \cos x + x^3 e^x \cos x - x^3 e^x \sin x.$

5. Derivative of a Quotient

If both u and v are differentiable, and $v(x) \neq 0$, their ratio is also differentiable:

$$\left(\frac{u}{v}\right)' = \frac{v\,u' - u\,v'}{v^2}, \quad d\left(\frac{u}{v}\right) = \frac{v\,du - u\,dv}{v^2}. \tag{6.8}$$

■ $y = \tan x = \dfrac{\sin x}{\cos x}, \quad y' = \dfrac{(\cos x)(\sin x)' - (\sin x)(\cos x)'}{\cos^2 x} = \dfrac{\cos^2 x + \sin^2 x}{\cos^2 x} = \dfrac{1}{\cos^2 x}.$

6. Chain Rule

The composite function (see 2.1.5.5, **2.**, p. 61) $y = u(v(x))$ has the derivative

$$\frac{dy}{dx} = u'(v)\,v'(x) = \frac{du}{dv}\frac{dv}{dx}, \tag{6.9}$$

where the functions $u = u(v)$ and $v = v(x)$ must be differentiable functions with respect to their own variables. $u(v)$ is called the exterior function, and $v(x)$ is called the interior function. According to this, $\dfrac{du}{dv}$ is the *exterior derivative* and $\dfrac{dv}{dx}$ is the *interior derivative*. It is possible that the functions u and v are not differentiable separately, but the composite function is. Then one gets the derivative by the definition formula (6.1).

Similarly one has to proceed if there is a longer "chain", i.e., in the case of a composite function of several *intermediate variables*. For example for $y = u(v(w(x)))$:

$$y' = \frac{dy}{dx} = \frac{du}{dv}\,\frac{dv}{dw}\,\frac{dw}{dx}. \tag{6.10}$$

■ **A:** $y = e^{\sin^2 x}, \quad \dfrac{dy}{dx} = \dfrac{d\left(e^{\sin^2 x}\right)}{d\left(\sin^2 x\right)}\,\dfrac{d\left(\sin^2 x\right)}{d\,(\sin x)}\,\dfrac{d\,(\sin x)}{dx} = e^{\sin^2 x}\,2\,\sin x \cos x.$

■ **B:** $y = e^{\tan\sqrt{x}}; \quad \dfrac{dy}{dx} = \dfrac{d\left(e^{\tan\sqrt{x}}\right)}{d\,(\tan\sqrt{x})}\,\dfrac{d\,(\tan\sqrt{x})}{d\,(\sqrt{x})}\,\dfrac{d(\sqrt{x})}{dx} = e^{\tan\sqrt{x}}\,\dfrac{1}{\cos^2\sqrt{x}}\,\dfrac{1}{2\sqrt{x}}.$

7. Logarithmic Differentiation

If $y(x) > 0$ holds, one can calculate the derivative y' starting with the function $\ln y(x)$, whose derivative (considering the chain rule) is

$$\frac{d(\ln y(x))}{dx} = \frac{1}{y(x)}y'. \tag{6.11}$$

From this rule

$$y' = y(x)\,\frac{d(\ln y(x))}{dx} \tag{6.12}$$

follows.

Remark 1: With the help of logarithmic differentiation it is possible to simplify some differentiation problems, and there are functions such that this is the only way to calculate the derivative, for instance, when the function has the form

$$y = u(x)^{v(x)} \quad \text{with} \quad u(x) > 0. \tag{6.13}$$

The logarithmic differentiation of this equality follows from the formula (6.12)

$$y' = y\,\frac{d\,(\ln u^v)}{dx} = y\,\frac{d(v \ln u)}{dx} = u^v\left(v' \ln u + \frac{vu'}{u}\right). \tag{6.14}$$

■ $y = (2x+1)^{3x}$, $\ln y = 3x\ln(2x+1)$, $\dfrac{y'}{y} = 3\ln(2x+1) + \dfrac{3x\cdot 2}{2x+1}$;
$y' = 3\,(2x+1)^{3x}\left(\ln(2x+1) + \dfrac{2x}{2x+1}\right)$.

Remark 2: Logarithmic differentiation is often used in the case to differentiate a product of several functions.

■ **A:** $y = \sqrt{x^3 e^{4x}\sin x}$, $\ln y = \dfrac{1}{2}(3\ln x + 4x + \ln\sin x)$,
$\dfrac{y'}{y} = \dfrac{1}{2}\left(\dfrac{3}{x} + 4 + \dfrac{\cos x}{\sin x}\right)$, $y' = \dfrac{1}{2}\sqrt{x^3 e^{4x}\sin x}\left(\dfrac{3}{x} + 4 + \cot x\right)$.

■ **B:** $y = u\,v$, $\ln y = \ln u + \ln v$, $\dfrac{y'}{y} = \dfrac{1}{u}u' + \dfrac{1}{v}v'$. From this identity it follows that $y' = (u\,v)' = v\,u' + u\,v'$, so one gets the formula for the derivative of a product (6.7a) (under the assumption $u, v > 0$).

■ **C:** $y = \dfrac{u}{v}$, $\ln y = \ln u - \ln v$, $\dfrac{y'}{y} = \dfrac{1}{u}u' - \dfrac{1}{v}v'$. From this identity it follows that $y' = \left(\dfrac{u}{v}\right)' = \dfrac{u'}{v} - \dfrac{uv'}{v^2} = \dfrac{v\,u' - u\,v'}{v^2}$, which is the formula for the derivative of a quotient (6.8) (under the assumption $u, v > 0$).

8. Derivative of the Inverse Function

If $y = \varphi(x)$ is the inverse function of the original function $y = f(x)$, then both forms $y = f(x)$ and $x = \varphi(y)$ are equivalent. For every corresponding value of x and y such that f is differentiable with respect to x, and φ is differentiable with respect to y, e.g., none of the derivatives is equal to zero, between the derivatives of f and its inverse function φ is valid the following relation:

$$f'(x) = \frac{1}{\varphi'(y)} \quad \text{or} \quad \frac{dy}{dx} = \frac{1}{\dfrac{dx}{dy}}. \tag{6.15}$$

■ The function $y = f(x) = \arcsin x$ for $-1 < x < 1$ is equivalent to the function $x = \varphi(y) = \sin y$ for $-\pi/2 < y < \pi/2$. From (6.15) it follows that
$(\arcsin x)' = \dfrac{1}{(\sin y)'} = \dfrac{1}{\cos y} = \dfrac{1}{\sqrt{1-\sin^2 y}} = \dfrac{1}{\sqrt{1-x^2}}$, because $\cos y \neq 0$ for $-\pi/2 < y < \pi/2$.

9. Derivative of an Implicit Function

Suppose the function $y = f(x)$ is given in implicit form by the equation $F(x, y) = 0$. Considering the rules of differentiation for functions of several variables (see 6.2, p. 445) calculating the derivative with respect to x gives

$$\frac{\partial F}{\partial x} + \frac{\partial F}{\partial y}y' = 0 \quad \text{and so} \quad y' = -\frac{F_x}{F_y}, \tag{6.16}$$

if the partial derivative F_y differs from zero.

■ The equation $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$ of an ellipse with semi-axes a and b can be written in the form $F(x, y) = \frac{x^2}{a^2} + \frac{y^2}{b^2} - 1 = 0$. For the slope of the tangent line at the point of the ellipse (x, y) one gets according to (6.16)

$$y' = -\frac{2x}{a^2}\Big/\frac{2y}{b^2} = -\frac{b^2}{a^2}\frac{x}{y}.$$

10. Derivative of a Function Given in Parametric Form

If a function $y = f(x)$ is given in parametric form $x = x(t)$, $y = y(t)$, then the derivative y' can be calculated by the formula

$$\frac{dy}{dx} = f'(x) = \frac{\dot{y}}{\dot{x}} \tag{6.17}$$

with the help of the derivatives $\dot{y}(t) = \frac{dy}{dt}$ and $\dot{x}(t) = \frac{dx}{dt}$ with respect to the variable t, if of course $\dot{x}(t) \neq 0$ holds.

■ **Polar Coordinate Representation:** If a function is given with polar coordinates (see 3.5.2.2, **3.**, p. 192) $\rho = \rho(\varphi)$, then the parametric form is

$$x = \rho(\varphi)\cos\varphi, \quad y = \rho(\varphi)\sin\varphi \tag{6.18}$$

with the angle φ as a parameter. For the slope y' of the tangent of the curve (see 3.6.1.2, **2.**, p. 245 or 6.1.1, **2.**, p. 432) one gets from (6.17)

$$y' = \frac{\dot{\rho}\sin\varphi + \rho\cos\varphi}{\dot{\rho}\cos\varphi - \rho\sin\varphi} \quad \text{where } \dot{\rho} = \frac{d\rho}{d\varphi}. \tag{6.19}$$

Remarks:

1. The derivatives $\dot{x}$, $\dot{y}$ are the components of the tangent vector at the point $(x(t), y(t))$ of the curve.

2. It is often useful to consider the complex relation:

$$x(t) + \mathrm{i}\,y(t) = z(t), \quad \dot{x}(t) + \mathrm{i}\,\dot{y}(t) = \dot{z}(t). \tag{6.20}$$

■ **Circular Movement:** $z(t) = re^{\mathrm{i}\omega t}$ $(r, \omega \text{ const})$, $\dot{z}(t) = r\mathrm{i}\omega e^{\mathrm{i}\omega t} = r\omega e^{\mathrm{i}(\omega t + \frac{\pi}{2})}$. The tangent vector runs ahead by a phase-shift $\pi/2$ with respect to the position vector.

11. Graphical Differentiation

If a differentiable function $y = f(x)$ is represented by its curve Γ in the Cartesian coordinate system in an interval $a < x < b$, then the curve Γ' of its derivative can be constructed approximately. The construction of a tangent estimated by eye is pretty inaccurate. However, if the direction of the tangent MN **(Fig. 6.4)** is given, then one can determine the point of contact A more precisely.

1. Construction of the Point of Contact of a Tangent

One draws two secants $\overline{M_1N_1}$ and $\overline{M_2N_2}$ parallel to the direction MN of the tangent so that the curve is intersected in points being not far from each other. Then there are to be determine the midpoints of the secants, and a straight line through them must be drawn. This line PQ intersects the curve at the point A, which is approximately the point, where the tangent has the given direction MN. To check the accuracy, one draws a third line close to and parallel to the first two lines, and the line PQ should intersect it at the midpoint.

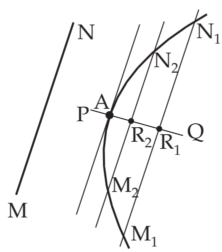


Figure 6.4

2. Construction of the Derivative Curve

a) Choose some directions $l_1, l_2, \ldots, l_n$ which could be the directions of some tangents of the curve

$y = f(x)$ in the considered interval as in **Fig. 6.5**, and determine the corresponding points of contact $A_1, A_2, \ldots, A_n$, where the tangents themselves must not be constructed.

b) Choose a point P, a "pole", on the negative side of the x-axis, where the longer the segment $PO = a$, the flatter the curve is.

c) Draw the lines through the pole P parallel to the directions $l_1, l_2, \ldots l_n$, and denote their intersection points with the y-axis by $B_1, B_2, \ldots B_n$.

d) Construct the horizontal lines $B_1C_1, B_2C_2, \ldots, B_nC_n$ through the points $B_1, B_2, \ldots, B_n$ to the intersection points $C_1, C_2, \ldots, C_n$ with the orthogonal lines from the points $A_1, A_2, \ldots, A_n$.

e) Connect the points $C_1, C_2, \ldots, C_n$ with the help of a curved ruler. The resulting curve satisfies the equation $y = af'(x)$. If the segment a is chosen so that it corresponds to the unit length on the y-axis, then the curve one gets is the curve of the derivative. Otherwise, one has to multiply the ordinates of $C_1, C_2, \ldots, C_n$ by the factor $1/a$. The points $D_1, D_2, \ldots, D_n$ given in **Fig. 6.5** are on the correctly scaled curve Γ' of the derivative.

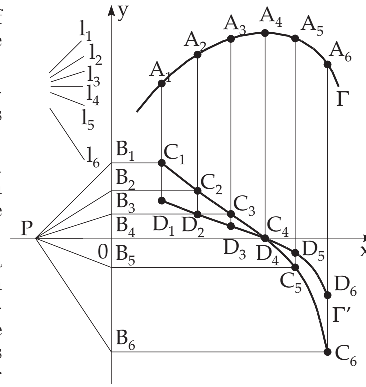


Figure 6.5

6.1.3 Derivatives of Higher Order

6.1.3.1 Definition of Derivatives of Higher Order

The derivative of $y' = f'(x)$, which means $(y')'$ or $\dfrac{d}{dx}\left(\dfrac{dy}{dx}\right)$, is called the second derivative of the function $y = f(x)$ and it is denoted by $y'', \ddot{y}, \dfrac{d^2y}{dx^2}, f''(x)$ or $\dfrac{d^2f(x)}{dx^2}$. Higher derivatives can be defined analogously. The notation for the n-th *derivative* of the function $y = f(x)$ is:

$$y^{(n)} = \frac{d^ny}{dx^n} = f^{(n)}(x) = \frac{d^nf(x)}{dx^n} \quad \left(n = 0, 1, \ldots;\ y^{(0)}(x) = f^{(0)}(x) = f(x)\right). \tag{6.21}$$

6.1.3.2 Derivatives of Higher Order of some Elementary Functions

The n-th derivatives of the simplest functions are collected in **Table 6.3**.

6.1.3.3 Leibniz's Formula

To calculate the n-th-order derivative of a product of two functions, the Leibniz formula can be used:

$$D^n(uv) = u\,D^nv + \frac{n}{1!}Du\,D^{n-1}v + \frac{n(n-1)}{2!}D^2u\,D^{n-2}v + \cdots + \frac{n(n-1)\ldots(n-m+1)}{m!}D^mu\,D^{n-m}v + \cdots + D^nu\,v \tag{6.22}$$

Here the notation $D^n = \dfrac{d^n}{dx^n}$ is used. If D^0u is replaced by u and D^0v by v, then one gets the formula (6.23) whose structure corresponds to the binomial formula (see 1.1.6.4, p. 12):

$$D^n(uv) = \sum_{m=0}^{n}\binom{n}{m}D^muD^{n-m}v\,. \tag{6.23}$$

Table 6.2 Differentiation rules

Expression	**Formula for the derivative**
Constant function	$c' = 0 \quad (c \text{ const})$
Constant multiple	$(cu)' = cu' \quad (c \text{ const})$
Sum	$(u \pm v)' = u' \pm v'$
Product of two functions	$(uv)' = u'v + uv'$
Product of n functions	$(u_1 u_2 \cdots u_n)' = \sum\limits_{i=1}^{n} u_1 \cdots u_i' \cdots u_n$
Quotient	$\left(\frac{u}{v}\right)' = \frac{vu' - uv'}{v^2} \quad (v \neq 0)$
Chain rule for two functions	$y = u(v(x)): \quad y' = \frac{du}{dv}\frac{dv}{dx}$
Chain rule for three functions	$y = u(v(w(x))): \quad y' = \frac{du}{dv}\frac{dv}{dw}\frac{dw}{dx}$
Power	$(u^\alpha)' = \alpha u^{\alpha-1}u' \quad (\alpha \in \mathbb{R},\ \alpha \neq 0)$ specially : $\left(\frac{1}{u}\right)' = -\frac{u'}{u^2} \quad (u \neq 0)$
Logarithmic differentiation	$\frac{d(\ln y(x))}{dx} = \frac{1}{y}y' \Longrightarrow y' = y\frac{d(\ln y)}{dx}$ special : $(u^v)' = u^v\left(v'\ln u + \frac{vu'}{u}\right) \quad (u > 0)$
Differentiation of the inverse function	φ inverse function of f, i.e. $y = f(x) \Longleftrightarrow x = \varphi(y)$: $f'(x) = \frac{1}{\varphi'(y)}$ or $\frac{dy}{dx} = \frac{1}{\frac{dx}{dy}}$
Implicit differentiation	$F(x,y) = 0: \quad F_x + F_y\, y' = 0$ or $y' = -\frac{F_x}{F_y} \quad \left(F_x = \frac{\partial F}{\partial x},\ F_y = \frac{\partial F}{\partial y};\ F_y \neq 0\right)$
Derivative in parameter form	$x = x(t),\ y = y(t)$ (t parameter): $y' = \frac{dy}{dx} = \frac{\dot{y}}{\dot{x}} \quad \left(\dot{x} = \frac{dx}{dt},\ \dot{y} = \frac{dy}{dt}\right)$
Derivative in polar coordinates	$r = r(\varphi): \quad \begin{array}{l} x = \rho(\varphi)\cos\varphi \\ y = \rho(\varphi)\sin\varphi \end{array}$ (angle φ as parameter) $y' = \frac{dy}{dx} = \frac{\dot{\rho}\sin\varphi + \rho\cos\varphi}{\dot{\rho}\cos\varphi - r\sin\varphi} \quad \left(\dot{\rho} = \frac{dx}{d\varphi}\right)$

■ **A:** $(x^2\cos ax)^{(50)}$: If $v = x^2,\ u = \cos ax$ are substituted, then follows $u^{(k)} = a^k\cos\left(ax + k\frac{\pi}{2}\right),\ v' = 2x,\ v'' = 2,\ v''' = v^{(4)} = \cdots = 0$. Except the first three cases, all the summands are equal to zero, so $(uv)^{(50)} = x^2a^{50}\cos\left(ax + 50\frac{\pi}{2}\right) + \frac{50}{1}\cdot 2xa^{49}\cos\left(ax + 49\frac{\pi}{2}\right) + \frac{50\cdot 49}{1\cdot 2}\cdot 2a^{48}\cos\left(ax + 48\frac{\pi}{2}\right)$

$= a^{48}[(2450 - a^2x^2)\cos ax - 100ax\sin ax]$.

■ **B:** $(x^3e^x)^{(6)} = \binom{6}{0} \cdot x^3e^x + \binom{6}{1} \cdot 3x^2e^x + \binom{6}{2} \cdot 6xe^x + \binom{6}{3} \cdot 6e^x = e^x(x^3 + 18x^2 + 90y + 120)$.

Table 6.3 Derivatives of higher order of some elementary functions

Function	***n*-th-order derivative**
x^m	$m(m-1)(m-2)\dots(m-n+1)x^{m-n}$ (for integer m and $n > m$ the n-th derivative is 0)
$\ln x \ (x > 0)$	$(-1)^{n-1}(n-1)!\,\frac{1}{x^n}$
$\log_a x \ (x > 0)$	$(-1)^{n-1}\frac{(n-1)!}{\ln a}\,\frac{1}{x^n}$
e^{kx}	$k^n e^{kx}$
a^x	$(\ln a)^n a^x$
a^{kx}	$(k\ln a)^n a^{kx}$
$\sin x$	$\sin(x + \frac{n\pi}{2})$
$\cos x$	$\cos(x + \frac{n\pi}{2})$
$\sin kx$	$k^n \sin(kx + \frac{n\pi}{2})$
$\cos kx$	$k^n \cos(kx + \frac{n\pi}{2})$
$\sinh x$	$\sinh x$ for even n, $\cosh x$ for odd n
$\cosh x$	$\cosh x$ for even n, $\sinh x$ for odd n

6.1.3.4 Higher Derivatives of Functions Given in Parametric Form

If a function $y = f(x)$ is given in the parametric form $x = x(t)$, $y = y(t)$, then its higher derivatives (y'', y''', etc.) can be calculated by the following formulas, where $\dot{y}(t) = \frac{dy}{dt}$, $\dot{x}(t) = \frac{dx}{dt}$, $\ddot{y}(t) = \frac{d^2y}{dt^2}$, $\ddot{x} = \frac{d^2x}{dt^2}$, etc., denote the derivatives with respect to the parameter t:

$$\frac{d^2y}{dx^2} = \frac{\dot{x}\ddot{y} - \dot{y}\ddot{x}}{\dot{x}^3}, \quad \frac{d^3y}{dx^3} = \frac{\dot{x}^2\dddot{y} - 3\dot{x}\ddot{x}\ddot{y} + 3\dot{y}\ddot{x}^2 - \dot{x}\dot{y}\dddot{x}}{\dot{x}^5}, \dots \quad (\dot{x}(t) \neq 0). \tag{6.24}$$

6.1.3.5 Derivatives of Higher Order of the Inverse Function

If $y = \varphi(x)$ is the inverse function of the original function $y = f(x)$, then both forms $y = f(x)$ and $x = \varphi(y)$ are equivalent. Supposing $\varphi'(y) \neq 0$ holds, the relation (6.15) is valid for the derivatives of the function f and its inverse function φ. For higher derivatives (y'', y''', etc.) one gets

$$\frac{d^2y}{dx^2} = -\frac{\varphi''(y)}{[\varphi'(y)]^3}, \quad \frac{d^3y}{dx^3} = \frac{3[\varphi''(y)]^2 - \varphi'(y)\varphi'''(y)}{[\varphi'(y)]^5}, \dots . \tag{6.25}$$

6.1.4 Fundamental Theorems of Differential Calculus

6.1.4.1 Monotonicity

If a function $f(x)$ is defined and continuous in a connected interval, and if it is differentiable at every interior point of this interval, then the relations

$$f'(x) \geq 0 \quad \text{for a monotone increasing function,} \tag{6.26a}$$

$$f'(x) \leq 0 \quad \text{for a monotone decreasing function} \tag{6.26b}$$

are necessary and sufficient. If the function is strictly monotone increasing or decreasing, then the derivative function $f'(x)$ must not be identically zero on any subinterval of the given interval. In **Fig. 6.6b** this condition is not fulfilled on the segment $\overline{BC}$.

The geometrical meaning of monotonicity is that the curve of an increasing function never falls for increasing values of the argument, i.e., it either rises or runs horizontally **(Fig. 6.6a)**. Therefore the tangent line at any point of the curve forms an acute angle with the positive x-axis or it is parallel to it. For monotonically decreasing functions **(Fig. 6.6b)** analogous statements are valid. If the function is strictly monotone, then the tangent can be parallel to the x-axis only at some single points, e.g., at the point A in **Fig. 6.6a**, i.e., not on a subinterval such as $\overline{BC}$ in **Fig. 6.6b**.

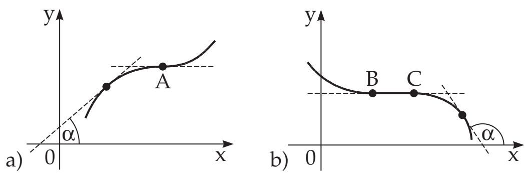


Figure 6.6

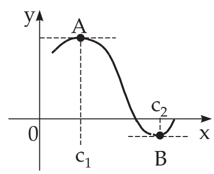


Figure 6.7

6.1.4.2 Fermat's Theorem

If a function $y = f(x)$ is defined on a connected interval, and it has a maximum or a minimum value at an interior point $x = c$ of this interval **(Fig. 6.7)**, i.e., if for every x in this interval

$f(c) > f(x)$ (6.27a) or $f(c) < f(x),$ (6.27b)

holds, and if the derivative exists at the point c, then the derivative must be equal to zero there:

$$f'(c) = 0. \tag{6.27c}$$

The geometrical meaning of the Fermat theorem is that if a function satisfies the assumptions of the theorem, then its curve has tangents parallel to the x-axis at A and B **(Fig. 6.7)**.

The Fermat theorem gives only a necessary condition for the existence of a maximum or minimum value at a point. From **Fig. 6.6a** it is obvious that having a zero derivative is not sufficient to give an extreme value: At the point A, $f'(x) = 0$ holds, but there is no maximum or minimum here.

To have an extreme value differentiability is not a necessary condition. The function in **Fig. 6.8d** has a maximum at e, but the derivative does not exist here.

6.1.4.3 Rolle's Theorem

If a function $y = f(x)$ is continuous on the closed interval $[a, b]$, and differentiable on the open interval (a, b), and

$$f(a) = 0, \quad f(b) = 0 \quad (a < b) \tag{6.28a}$$

hold, then there exists at least one point c between a and b such that

$$f'(c) = 0 \quad (a < c < b) \tag{6.28b}$$

holds. The geometrical meaning of Rolle's theorem is that if the graph of a function $y = f(x)$ which is continuous on the interval (a, b) intersects the x-axis at two points A and B, and it has a non-vertical

tangent at every point, then there is at least one point C between A and B such that the tangent is parallel to the x-axis here **(Fig. 6.8a)**. It is possible, that there are several such points in this inter-

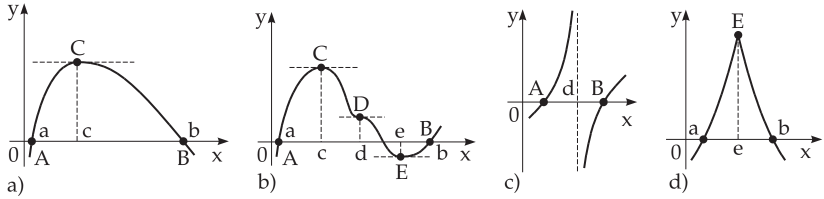


Figure 6.8

val, e.g., the points C, D, and E in **Fig. 6.8b**. The properties of continuity and differentiability are important in the theorem: in **Fig. 6.8c** the function is not continuous at $x = d$, and in **Fig. 6.8d** the function is not differentiable at $x = e$. In both cases $f'(x) \neq 0$ holds everywhere where the derivative exists.

6.1.4.4 Mean Value Theorem of Differential Calculus

If a function $y = f(x)$ is continuous on the closed interval $[a, b]$ and differentiable on the open interval (a, b), then there is at least one point c between a and b which satisfies the following relation:

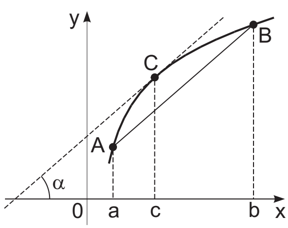


Figure 6.9

$$\frac{f(b) - f(a)}{b - a} = f'(c) \quad (a < c < b) \tag{6.29a}$$

holds. Substituting $b = a + h$, and Θ means a number between 0 and 1, then the theorem can be written in the form

$$f(a + h) = f(a) + h\, f'(a + \Theta\, h) \quad (0 < \Theta < 1). \tag{6.29b}$$

1. Geometrical Meaning The geometrical meaning of the theorem is that if a function $y = f(x)$ satisfies the conditions of the theorem, then its graph has at least one point C between A and B such that the tangent line at this point is parallel to the line segment between A and B **(Fig. 6.9)**. There can be several such points **(Fig. 6.8b)**.

That the properties of continuity and differentiability are important can be shown in examples and also as can be observed in **Fig. 6.8c,d**.

2. Applications The mean value theorem has several useful applications.

■ **A:** This theorem can be used to prove some inequalities in the form

$$|f(b) - f(a)| < K|b - a|, \tag{6.30}$$

where K is an upper bound of $|f'(x)|$ for every x in the interval $[a, b]$.

■ **B:** How accurate is the value of $f(\pi) = \dfrac{1}{1 + \pi^2}$ if π is replaced by the approximate value $\overline{\pi} = 3.14$?

We have: $|f(\pi) - f(\overline{\pi})| = \left|\dfrac{2c}{(1 + c^2)^2}\right| |\pi - \overline{\pi}| \leq 0.053 \cdot 0.0016 = 0.000\,085$, which means $\dfrac{1}{1 + \pi^2}$ is between $0.092\,084 \pm 0.000\,085$.

6.1.4.5 Taylor's Theorem of Functions of One Variable

If a function $y = f(x)$ is continuously differentiable (it has continuous derivatives) $n - 1$ times on the interval $[a, a + h]$, and if also the n-th derivative exists in the interior of the interval, then the Taylor

formula or Taylor expansion is

$$f(a+h) = f(a) + \frac{h}{1!}f'(a) + \frac{h^2}{2!}f''(a) + \cdots + \frac{h^{n-1}}{(n-1)!}f^{(n-1)}(a) + \frac{h^n}{n!}f^{(n)}(a+\Theta h) \qquad (6.31)$$

with $0 < \Theta < 1$. The quantity h can be positive or negative. The mean value theorem (6.29b) is a special case of the Taylor formula for $n = 1$.

6.1.4.6 Generalized Mean Value Theorem of Differential Calculus (Cauchy's Theorem)

If two functions $y = f(x)$ and $y = \varphi(x)$ are continuous on the closed interval $[a, b]$ and they are differentiable at least in the interior of the interval, and $\varphi'(x)$ is never equal to zero in this interval, then there exists at least one value c between a and b such that

$$\frac{f(b) - f(a)}{\varphi(b) - \varphi(a)} = \frac{f'(c)}{\varphi'(c)} \quad (a < c < b). \qquad (6.32)$$

The geometrical meaning of the generalized mean value theorem corresponds to that of the first mean value theorem. Supposing, e.g., that the curve in **Fig.6.9** is given in parametric form $x = \varphi(t)$, $y = f(t)$, where the points A and B belong to the parameter values $t = a$ and $t = b$ respectively. Then for the point C

$$\tan\alpha = \frac{f(b) - f(a)}{\varphi(b) - \varphi(a)} = \frac{f'(c)}{\varphi'(c)} \qquad (6.33)$$

is valid. For $\varphi(x) = x$ the generalized mean value theorem is simplified into the first mean value theorem.

6.1.5 Determination of the Extreme Values and Inflection Points

6.1.5.1 Maxima and Minima

The substitution value $f(x_0)$ of a function $f(x)$ is called the *relative maximum* (M) or *relative minimum* (m) if one of the inequalities

$$f(x_0 + h) < f(x_0) \quad \text{(for maximum)}, \qquad (6.34a)$$

$$f(x_0 + h) > f(x_0) \quad \text{(for minimum)} \qquad (6.34b)$$

holds for arbitrary positive or negative values of h small enough. At a relative maximum the value $f(x_0)$ is greater than the values in the neighborhood, and similarly, at a minimum it is smaller. The relative maxima and minima are called *relative* or *local extrema*. The greatest or the smallest value of a function in an interval is called the *global* or *absolute maximum* or *global* or *absolute minimum* in this interval.

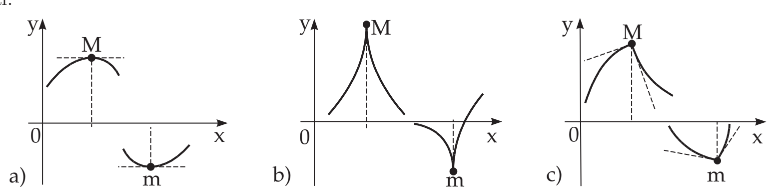


Figure 6.10

6.1.5.2 Necessary Conditions for the Existence of a Relative Extreme Value

A function can have a relative maximum or minimum only at the points where its derivative is equal to zero or does not exist. That is: At the points of the graph of the function corresponding to the relative extrema the tangent line is whether parallel to the x-axis **(Fig. 6.10a)** or parallel to the y-axis

(Fig. 6.10b) or does not exist **(Fig. 6.10c)**. Anyway, these are not sufficient conditions, e.g., at the points A, B, C in **Fig. 6.11** these conditions are obviously fulfilled, but there are no extreme values of the function.
If a continuous function has relative extreme values, then maxima and minima follow alternately, that means, between two neighboring maxima there is a minimum, and conversely.

6.1.5.3 Determination of the Relative Extreme Values and the Inflection Points of a Differentiable Explicit Function $y = f(x)$

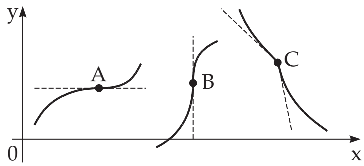


Figure 6.11

Since $f'(x) = 0$ is a necessary condition where the derivative exists, after determining the derivative $f'(x)$, first one calculates all the real roots $x_1, x_2, \ldots, x_i, \ldots, x_n$ of the equation $f'(x) = 0$. Then each of them has to be checked , e.g., x_i with one of the following methods.

1. Method of Sign Change

For values x_- and x_+, which are slightly smaller and greater than x_i, and for which between x_i and x_- and x_+ no more roots or points of discontinuity of $f'(x)$ exist, one checks the sign of $f'(x)$. When during the transition from $f'(x_-)$ to $f'(x_+)$ the sign of $f'(x)$ changes from "+" to "−", then there is a relative maximum of the function $f(x)$ at $x = x_i$ **(Fig. 6.12a)**; if it changes from "−" to "+", then there is a relative minimum there **(Fig. 6.12b)**. If the derivative does not change its sign **(Fig. 6.12c,d)**, then there is no extremum at $x = x_i$, but it has an inflection point with a tangent parallel to the x-axis.

2. Method of Higher Derivatives

If a function has higher derivatives at $x = x_i$, then one can substitute, e.g., the root x_i into the second derivative $f''(x)$. If $f''(x_i) < 0$ holds, then there is a relative maximum at x_i, and if $f''(x_i) > 0$ holds, a relative minimum. If $f''(x_i) = 0$ holds, then x_i must be substituted into the third derivative $f'''(x)$. If $f'''(x_i) \neq 0$ holds, then there is no extremum at $x = x_i$ but an inflection point. If still $f'''(x_i) = 0$ holds, then one substitutes it into the forth derivative, etc. If the first non-zero derivative at $x = x_i$ is an even one, then $f(x)$ has an extremum here: If the derivative is positive, then there is minimum, if it is negative, then there is a maximum. If the first non-zero derivative is an odd one, then there is no extremum there (actually, there is an inflection point).

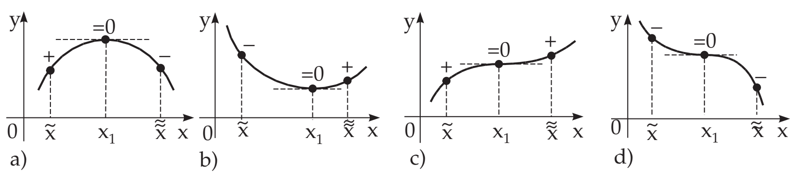


Figure 6.12

3. Further Conditions for Extreme Points and Determination of Inflection Points

If a continuous function is increasing below x_0 and decreasing after, then it has a maximum there; if it is decreasing below and increasing after, then it has a minimum there. Checking the sign change of the derivative is a useful method even if the derivative does not exist at certain points as in **Fig. 6.10b,c** and **Fig. 6.11**. If the first derivative exists at a point where the function has an inflection point, then the first derivative has an extremum there. So, to find the inflection points with the help of derivatives, one has to do the same investigation for the derivative function as one has done for the original function to find its extrema.

Remark: For non-continuous functions, and sometimes also for certain differentiable functions the

determination of extrema needs individual ideas. It is possible that a function has an extremum so that the first derivative exists and it is equal to zero, but the second derivative does not exist, and the first one has infinitely many roots in an arbitrary neighborhood of the considered point, so it is meaningless to say it changes its sign there. For instance $f(x) = x^2(2 + \sin(1/x))$ for $x \neq 0$ and $f(0) = 0$.

6.1.5.4 Determination of Absolute Extrema

The considered interval of the independent variable is divided into subintervals such that in these intervals the function has a continuous derivative. The absolute extreme values are among the relative extreme values, or at the endpoints of the subintervals, if their endpoints belong to them. For non-continuous functions or for non-closed intervals it is possible that no maximum or minimum exists on the considered interval.

Examples of the Determination of Extrema:

■ **A:** $y = e^{-x^2}$, interval $[-1, +1]$. Greatest value at $x = 0$, smallest at the endpoints **(Fig. 6.13a).**

■ **B:** $y = x^3 - x^2$, interval $[-1, +2]$. Greatest value at $x = +2$, smallest at $x = -1$, at the ends of the interval **(Fig. 6.13b)**.

■ **C:** $y = \dfrac{1}{1 + e^{\frac{1}{x}}}$, interval $[-3, +3]$, $x \neq 0$. There is no maximum or minimum. Relative minimum at $x = -3$, relative maximum at $x = 3$. If one defines $y = 1$ for $x = 0$, then there will be an absolute maximum at $x = 0$ **(Fig. 6.13c).**

■ **D:** $y = 2 - x^{\frac{2}{3}}$, interval $[-1, +1]$. Greatest value at $x = 0$ (**Fig. 6.13d**, the derivative is not finite).

6.1.5.5 Determination of the Extrema of Implicit Functions

If the function is given in the implicit form $F(x, y) = 0$, and the function F itself and also its partial derivatives F_x, F_y are continuous, then its maxima and minima can be determined in the following way:

1. Solution of the Equation System $F(x, y) = 0, F_x(x, y) = 0$ and substitution of the resulting values $(x_1, y_1), (x_2, y_2), \ldots, (x_i, y_i), \ldots$ in F_y and F_{xx}.

2. Sign Comparison for F_y and F_{xx} at the Point (x_i, y_i): When they have different signs, the function $y = f(x)$ has a minimum at x_i; when F_y and F_{xx} have the same sign, then it has a maximum at x_i. If either F_y or F_{xx} vanishes at (x_i, y_i), then one needs further and rather complicated investigation.

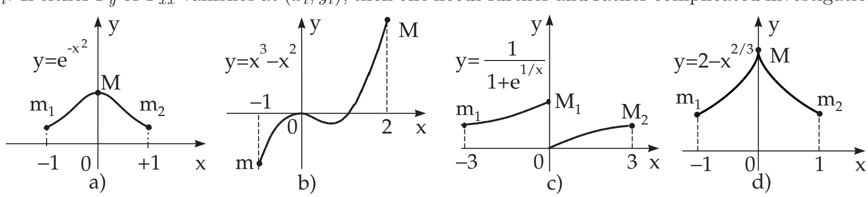


Figure 6.13

6.2 Differentiation of Functions of Several Variables

6.2.1 Partial Derivatives

6.2.1.1 Partial Derivative of a Function

The partial derivative of a function $u = f(x_1, x_2, \ldots, x_i, \ldots, x_n)$ with respect to one of its n variables, e.g., with respect to x_1 is defined by

$$\frac{\partial u}{\partial x_1} = \lim_{\Delta x_1 \to 0} \frac{f(x_1 + \Delta x_1, x_2, x_3, \ldots, x_n) - f(x_1, x_2, x_3, \ldots, x_n)}{\Delta x_1}, \tag{6.35}$$

so only one of the n variables is changing, the other $n-1$ are considered as constants. The symbols for the partial derivatives are $\frac{\partial u}{\partial x}$, u'_x, $\frac{\partial f}{\partial x}$, f'_x. A function of n variables can have n first-order partial derivatives: $\frac{\partial u}{\partial x_1}, \frac{\partial u}{\partial x_2}, \frac{\partial u}{\partial x_3}, \ldots, \frac{\partial u}{\partial x_n}$. The calculation of the partial derivatives can be done following the same rules as there are for the functions of one variable.

■ $u = \frac{x^2 y}{z}, \quad \frac{\partial u}{\partial x} = \frac{2xy}{z}, \quad \frac{\partial u}{\partial y} = \frac{x^2}{z}, \quad \frac{\partial u}{\partial z} = -\frac{x^2 y}{z^2}.$

6.2.1.2 Geometrical Meaning for Functions of Two Variables

If a function $u = f(x, y)$ is represented as a surface in a Cartesian coordinate system, and this surface is intersected through its point P by a plane parallel to the x, u plane **(Fig. 6.14)**, then holds

$$\frac{\partial u}{\partial x} = \tan\alpha, \tag{6.36a}$$

where α is the angle between the positive x-axis and the tangent line of the intersection curve at P, which is the same as the angle between the positive x-axis and the perpendicular projection of the tangent line into the x, u plane. Here, α is measured starting at the x-axis, and the positive direction is counterclockwise if looking toward the positive half of the y-axis. Analogously to α, β is defined with a plane parallel to the y, u plane:

$$\frac{\partial u}{\partial y} = \tan\beta. \tag{6.36b}$$

The derivative with respect to a given direction, the so-called *directional derivative*, and *derivative with respect to volume*, will be discussed in vector analysis (see 13.2.1, p. 708 and p. 709).

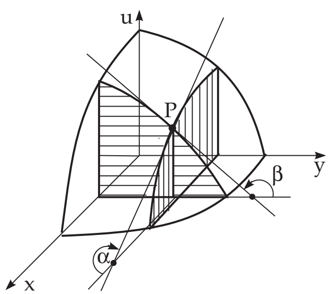


Figure 6.14

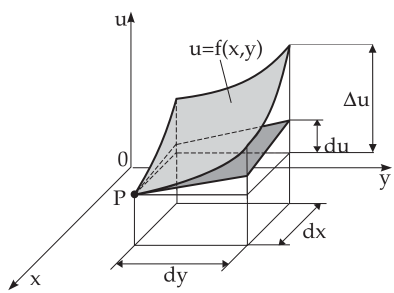


Figure 6.15

6.2.1.3 Differentials of x and $f(x)$

1. The Differential dx of an Independent Variable x

is equal to the increment Δx, i.e.,

$$dx = \Delta x \tag{6.37a}$$

for an arbitrary value of Δx.

2. The Differential dy of a Function $y = f(x)$ of One Variable x

is defined for a given value of x and for a given value of the differential dx as the product

$$dy = f'(x)\,dx. \tag{6.37b}$$

3. The Increment of a Function $y = f(x)$ for $x + \Delta x$
is the difference

$$\Delta y = f(x + \Delta x) - f(x). \tag{6.37c}$$

4. Geometrical Meaning of the Differential
If the function is represented by a curve in a Cartesian coordinate system, then dy is the increment of the ordinate of the tangent line for the change of x by a given increment dx **(Fig. 6.1)**. In an analogous way Δy is the increment of the ordinate of the curve.

6.2.1.4 Basic Properties of the Differential

1. Invariance
Independently of whether x is an independent variable or a function of a further variable t

$$dy = f'(x)\,dx \tag{6.38}$$

is valid.

2. Order of Magnitude
If dx is an arbitrarily small value, then dy and $\Delta y = y(x + \Delta x) - y(x)$ are also arbitrarily small, but of equivalent amounts, i.e., $\lim\limits_{\Delta x \to 0} \dfrac{\Delta y}{dy} = 1$. Consequently, the difference between them is also arbitrarily small, but of higher order than dx, dy and Δx (except if $dy = 0$ holds). Therefore, one gets the relation

$$\lim_{\Delta x \to 0} \frac{\Delta y}{dy} = 1, \quad \Delta y \approx dy = f'(x)\,dx, \tag{6.39}$$

which allows to reduce the calculation of a small *increment* to the calculation of its differential. This formula is frequently used for approximate calculations (see 6.1.4.4, p. 442 and 16.4.2.1, **2.**, p. 855).

6.2.1.5 Partial Differential

For a function of several variables $u = f(x, y, \ldots)$ one can form the partial differential with respect to one of its variables, e.g., with respect to x, which is defined by the equality

$$d_x u = d_x f = \frac{\partial u}{\partial x}\,dx. \tag{6.40}$$

6.2.2 Total Differential and Differentials of Higher Order

6.2.2.1 Notion of Total Differential of a Function of Several Variables (Complete Differential)

1. Differentiability
The function of several variable $u = f(x_1, x_2, \ldots, x_i, \ldots, x_n)$ is called *differentiable* at the point $P_0(x_{10}, x_{20}, \ldots, x_{i0}, \ldots, x_{n0})$ if at a transition to an arbitrarily close point $P(x_{10}+\Delta x_1, x_{20}+\Delta x_2, \ldots, x_{i0} +\Delta x_i, \ldots, x_{n0} + \Delta x_n)$ with the arbitrarily small quantities $\Delta x_1, \Delta x_2, \ldots, \Delta x_i, \ldots, \Delta x_n$ the complete increment

$$\begin{aligned}\Delta u = {} & f(x_{10} + \Delta x_1, x_{20} + \Delta x_2, \ldots, x_{i0} + \Delta x_i, \ldots, x_{n0} + \Delta x_n) \\ & - f(x_{10}, x_{20}, \ldots, x_{i0}, \ldots, x_{n0})\end{aligned} \tag{6.41a}$$

of the function differs from the sum of the partial differentials of all variables

$$\left(\frac{\partial u}{\partial x_1}dx_1 + \frac{\partial u}{\partial x_2}dx_2 + \ldots + \frac{\partial u}{\partial x_n}dx_n\right)_{x_{10}, x_{20}, \ldots, x_{n0}} \tag{6.41b}$$

by an arbitrarily small amount in higher order than the distance

$$\overline{P_0 P} = \sqrt{\Delta x_1^2 + \Delta x_2^2 + \ldots + \Delta x_n^2} = \sqrt{dx_1^2 + dx_2^2 + \ldots + dx_n^2}\,. \tag{6.41c}$$

A continuous function of several variables is differentiable at a point if its partial derivatives, as functions of several variables, are continuous in a neighborhood of this point. This is a sufficient but not a

necessary condition, while the simple existence of the partial derivatives at the considered point is not sufficient even for the continuity of the function.

2. Total Differential

If u is a differentiable function, then the sum (6.41b)

$$du = \frac{\partial u}{\partial x_1}dx_1 + \frac{\partial u}{\partial x_2}dx_2 + \ldots + \frac{\partial u}{\partial x_n}dx_n \tag{6.42a}$$

is called the *total differential* of the function. With the n-dimensional vectors

$$\underline{\mathbf{grad\,u}} = \left(\frac{\partial u}{\partial x_1}, \frac{\partial u}{\partial x_2}, \ldots, \frac{\partial u}{\partial x_n}\right)^{\mathrm{T}}, \quad (6.42b) \qquad \underline{\mathbf{dr}} = (dx_1, dx_2, \ldots, dx_n)^{\mathrm{T}} \tag{6.42c}$$

the total differential can be expressed as the scalar product

$$du = (\underline{\mathbf{grad\,u}})^{\mathrm{T}} \cdot \underline{\mathbf{dr}}. \tag{6.42d}$$

In (6.42b), there is the gradient, defined in 13.2.2, p. 710, for n independent variables.

3. Geometrical Representation

The geometrical meaning of the total differential of a function of two variables $u = f(x, y)$, represented in a Cartesian coordinate system as a surface **(Fig. 6.15)**, is that du is the same as the increment of the applicate (see 3.5.3.1, **3.**, p. 210) of the tangent plane (at the same point) if dx and dy are the increments of x and y.

From the Taylor formula (see 6.2.2.3, **1.**, p. 449) it follows for functions of two variables that

$$f(x, y) = f(x_0, y_0) + \frac{\partial f}{\partial x}(x_0, y_0)(x - x_0) + \frac{\partial f}{\partial y}(x_0, y_0)(y - y_0) + R_1. \tag{6.43a}$$

Ignoring the remainder R_1, holds that

$$u = f(x_0, y_0) + \frac{\partial f}{\partial x}(x_0, y_0)(x - x_0) + \frac{\partial f}{\partial y}(x_0, y_0)(y - y_0) \tag{6.43b}$$

gives the equation of the tangent plane of the surface $u = f(x, y)$ at the point $P_0(x_0, y_0, u_0)$.

4. The Fundamental Property of the Total Differential

is the invariance with respect to the variables as formulated in (6.38) for the one-variable case.

5. Application in Error Calculations

In error calculations one uses the total differential du for an estimation of the error Δu (see (6.41a)) (see, e.g., 16.4.1.3, **5.**, p. 852). From the Taylor formula (see 6.2.2.3, **1.**, p. 449) follows

$$|\Delta u| = |du + R_1| \le |du| + |R_1| \approx |du|, \tag{6.44}$$

i.e., the absolute error $|\Delta u|$ can be replaced by $|du|$ as a first approximation. It follows that du is a linear approximation for Δu.

6.2.2.2 Derivatives and Differentials of Higher Order

1. Partial Derivatives of Second Order, Schwarz's Exchange Theorem

The second-order partial derivative of a function $u = f(x_1, x_2, \ldots, x_i, \ldots, x_n)$ can be calculated with respect to the same variable as the first one was, i.e., $\frac{\partial^2 u}{\partial x_1^2}, \frac{\partial^2 u}{\partial x_2^2}, \ldots$, or with respect to another variable, i.e. $\frac{\partial^2 u}{\partial x_1 \partial x_2}, \frac{\partial^2 u}{\partial x_2 \partial x_3}, \frac{\partial^2 u}{\partial x_3 \partial x_1}, \ldots$. In this second case one talks about mixed derivatives.

If at the considered point the mixed derivatives are continuous, then

$$\frac{\partial^2 u}{\partial x_1 \partial x_2} = \frac{\partial^2 u}{\partial x_2 \partial x_1} \tag{6.45}$$

holds for given x_1 and x_2 independently of the order of sequence of the differentiation (*Schwarz's exchange theorem*).

Partial derivatives of higher order such as, e.g., $\dfrac{\partial^3 u}{\partial x^3}$, $\dfrac{\partial^3 u}{\partial x \partial y^2}$, ... are defined analogously.

2. Second-Order Differential of a Function of One Variable $y = f(x)$

The second-order differential of a function $y = f(x)$ of one variable, denoted by the symbols d^2y, $d^2f(x)$, is the differential of the first differential: $d^2y = d(dy) = f''(x)dx^2$. These symbols are appropriate only if x is an independent variable, and they are not appropriate if x is given, e.g., in the form $x = z(v)$. Differentials of higher order are defined analogously. If the variables $x_1, x_2, \ldots, x_i, \ldots, x_n$ are themselves functions of other variables, then one gets more complicated formulas (see 6.2.4, p. 452).

3. Total Differential of Second Order of a Function of Two Variables $u = f(x, y)$

$$d^2u = d(du) = \frac{\partial^2 u}{\partial x^2}dx^2 + 2\frac{\partial^2 u}{\partial x \partial y}\,dx\,dy + \frac{\partial^2 u}{\partial y^2}\,dy^2 \tag{6.46a}$$

or symbolically

$$d^2u = \left(\frac{\partial}{\partial x}\,dx + \frac{\partial}{\partial y}dy\right)^2 u. \tag{6.46b}$$

4. Total Differential of n-th Order of a Function of Two Variables

$$d^n u = \left(\frac{\partial}{\partial x}dx + \frac{\partial}{\partial y}dy\right)^n u. \tag{6.47}$$

5. Total Differential of n-th Order of a Function $u = f(x_1, x_2, \ldots, x_m)$ of m Variables

$$d^n u = \left(\frac{\partial}{\partial x_1}\,dx_1 + \frac{\partial}{\partial x_2}\,dx_2 + \ldots + \frac{\partial}{\partial x_m}\,dx_m\right)^n u. \tag{6.48}$$

6.2.2.3 Taylor's Theorem for Functions of Several Variables

1. Taylor's Formula for Functions of Two Variables

a) First Form of Representation:

$$\begin{aligned} f(x,y) &= f(a,b) + \left.\frac{\partial f(x,y)}{\partial x}\right|_{(x,y)=(a,b)}(x-a) + \left.\frac{\partial f(x,y)}{\partial y}\right|_{(x,y)=(a,b)}(y-b) \\ &+\frac{1}{2!}\left\{\left.\frac{\partial^2 f(x,y)}{\partial x^2}\right|_{(x,y)=(a,b)}(x-a)^2 + 2\left.\frac{\partial^2 f(x,y)}{\partial x \partial y}\right|_{(x,y)=(a,b)}(x-a)(y-b)\right. \\ &\left.+\left.\frac{\partial^2 f(x,y)}{\partial y^2}\right|_{(x,y)=(a,b)}(y-b)^2\right\} + \frac{1}{3!}\{\ldots\} + \cdots + \frac{1}{n!}\{\ldots\} + R_n. \end{aligned} \tag{6.49a}$$

Here (a, b) is the center of expansion and R_n is the remainder. Sometimes one writes, e.g., instead of $\left.\dfrac{\partial f(x,y)}{\partial x}\right|_{(x,y)=(x_0,y_0)}$ the shorter expression $\dfrac{\partial f}{\partial x}(x_0, y_0)$.

The terms of higher order in (6.49a) can be represented in a clear way with the help of operators:

$$\begin{aligned} f(x,y) &= f(a,b) + \frac{1}{1!}\left\{(x-a)\frac{\partial}{\partial x} + (y-b)\frac{\partial}{\partial y}\right\}f(x,y)\Big|_{(x,y)=(a,b)} \\ &+\frac{1}{2!}\left\{(x-a)\frac{\partial}{\partial x} + (y-b)\frac{\partial}{\partial y}\right\}^2 f(x,y)\Big|_{(x,y)=(a,b)} \end{aligned}$$

$$+\frac{1}{3!}\{\ldots\}^3 f(x,y)\Big|_{(x,y)=(a,b)} + \cdots + \frac{1}{n!}\{\ldots\}^n f(x,y)\Big|_{(x,y)=(a,b)} + R_n\,. \tag{6.49b}$$

This symbolic form means that after using the binomial theorem the powers of the differential operators $\frac{\partial}{\partial x}$ and $\frac{\partial}{\partial y}$ represent the higher-order derivatives of the function $f(x, y)$. Then the derivatives must be taken at the point (a, b).

b) Second Form of the Representation:

$$f(x+h, y+k) = f(x,y) + \frac{1}{1!}\left(\frac{\partial}{\partial x}h + \frac{\partial}{\partial y}k\right) f(x,y) + \frac{1}{2!}\left(\frac{\partial}{\partial x}h + \frac{\partial}{\partial y}k\right)^2 f(x,y)$$
$$+\frac{1}{3!}\left(\frac{\partial}{\partial x}h + \frac{\partial}{\partial y}k\right)^3 f(x,y) + \cdots + \frac{1}{n!}\left(\frac{\partial}{\partial x}h + \frac{\partial}{\partial y}k\right)^n f(x,y) + R_n. \tag{6.49c}$$

c) Remainder: The expression for the remainder is

$$R_n = \frac{1}{(n+1)!}\left(\frac{\partial}{\partial x}h + \frac{\partial}{\partial y}k\right)^{n+1} f(x+\Theta h, y+\Theta k) \qquad (0 < \Theta < 1). \tag{6.49d}$$

2. Taylor Formula for Functions of m Variables

The analogous representation with differential operators is

$$f(x+h,\ y\ +k, \ldots, t+l)$$
$$= f(x,y,\ldots,t) + \sum_{i=1}^{n}\frac{1}{i!}\left(\frac{\partial}{\partial x}h + \frac{\partial}{\partial y}k + \cdots + \frac{\partial}{\partial t}l\right)^i f(x,y,\ldots,t) + R_n, \tag{6.50a}$$

where the remainder can be calculated by the expression

$$R_n = \frac{1}{(n+1)!}\left(\frac{\partial}{\partial x}h + \frac{\partial}{\partial y}k + \cdots + \frac{\partial}{\partial t}l\right)^{n+1} f(x+\Theta\, h, y+\Theta\, k, \ldots, t+\Theta\, l)$$
$$(0 < \Theta < 1). \tag{6.50b}$$

6.2.3 Rules of Differentiation for Functions of Several Variables

6.2.3.1 Differentiation of Composite Functions

1. Composite Function of One Independent Variable

$$u = f(x_1, x_2, \ldots, x_n), \quad x_1 = x_1(\xi), \quad x_2 = x_2(\xi), \ldots, \quad x_n = x_n(\xi) \tag{6.51a}$$

$$\frac{\partial u}{\partial \xi} = \frac{\partial u}{\partial x_1}\frac{dx_1}{d\xi} + \frac{\partial u}{\partial x_2}\frac{dx_2}{d\xi} + \ldots + \frac{\partial u}{\partial x_n}\frac{dx_n}{d\xi}\,. \tag{6.51b}$$

2. Composite Function of Several Independent Variables

$$u = f(x_1, x_2, \ldots, x_n),$$
$$x_1 = x_1(\xi, \eta, \ldots, \tau), \quad x_2 = x_2(\xi, \eta, \ldots, \tau), \ldots, \quad x_n = x_n(\xi, \eta, \ldots, \tau) \tag{6.52a}$$

$$\left.\begin{aligned}
\frac{\partial u}{\partial \xi} &= \frac{\partial u}{\partial x_1}\frac{\partial x_1}{\partial \xi} + \frac{\partial u}{\partial x_2}\frac{\partial x_2}{\partial \xi} + \cdots + \frac{\partial u}{\partial x_n}\frac{\partial x_n}{\partial \xi},\\
\frac{\partial u}{\partial \eta} &= \frac{\partial u}{\partial x_1}\frac{\partial x_1}{\partial \eta} + \frac{\partial u}{\partial x_2}\frac{\partial x_2}{\partial \eta} + \cdots + \frac{\partial u}{\partial x_n}\frac{\partial x_n}{\partial \eta},\\
\vdots\ &= \quad \vdots \quad + \quad \vdots \quad + \vdots + \quad \vdots\\
\frac{\partial u}{\partial \tau} &= \frac{\partial u}{\partial x_1}\frac{\partial x_1}{\partial \tau} + \frac{\partial u}{\partial x_2}\frac{\partial x_2}{\partial \tau} + \cdots + \frac{\partial u}{\partial x_n}\frac{\partial x_n}{\partial \tau}.
\end{aligned}\right\} \tag{6.52b}$$

6.2.3.2 Differentiation of Implicit Functions

1. A Function $y = f(x)$ of One Variable is given by the equation

$$F(x,y) = 0. \tag{6.53a}$$

Differentiating (6.53a) with respect to x with the help of (6.51b) one gets

$$F_x + F_y y' = 0 \tag{6.53b}$$ and $$y' = -\frac{F_x}{F_y} \quad (F_y \neq 0). \tag{6.53c}$$

Differentiation of (6.53b) yields in the same way

$$F_{xx} + 2F_{xy}y' + F_{yy}(y')^2 + F_y y'' = 0, \tag{6.53d}$$

so considering (6.53b) one has

$$y'' = \frac{2F_xF_yF_{xy} - (F_y)^2F_{xx} - (F_x)^2F_{yy}}{(F_y)^3}. \tag{6.53e}$$

In an analogous way one can calculate the third derivative

$$F_{xxx} + 3F_{xxy}y' + 3F_{xyy}(y')^2 + F_{yyy}(y')^3 + 3F_{xy}y'' + 3F_{yy}y'y'' + F_y y''' = 0, \tag{6.53f}$$

from which y''' can be expressed.

2. A Function $u = f(x_1, x_2, \dots, x_i, \dots, x_n)$ of Several Variables is given by the equation

$$F(x_1, x_2, \dots, x_i, \dots, x_n, u) = 0. \tag{6.54a}$$

The partial derivatives

$$\frac{\partial u}{\partial x_1} = -\frac{F_{x_1}}{F_u}, \frac{\partial u}{\partial x_2} = -\frac{F_{x_2}}{F_u}, \dots, \frac{\partial u}{\partial x_n} = -\frac{F_{x_n}}{F_u} \tag{6.54b}$$

can be calculated similarly as it has been shown above but here the formulas (6.52b) are to be used. The higher-order derivatives can be calculated in the same way.

3. Two Functions $y = f(x)$ and $z = \varphi(x)$ of One Variable are given by the system of equations

$$F(x,y,z) = 0 \quad \text{and} \quad \Phi(x,y,z) = 0. \tag{6.55a}$$

Then differentiation of (6.55a) according to (6.51b) results in

$$F_x + F_y y' + F_z z' = 0, \quad \Phi_x + \Phi_y y' + \Phi_z z' = 0, \tag{6.55b}$$

$$y' = \frac{F_z\Phi_x - \Phi_zF_x}{F_y\Phi_z - F_z\Phi_y}, \quad z' = \frac{F_x\Phi_y - F_y\Phi_x}{F_y\Phi_z - F_z\Phi_y}. \tag{6.55c}$$

The second derivatives y'' and z'' are calculated in the same way by differentiation of (6.55b) considering y' and z'.

4. n Functions of One Variable Let the functions $y_1 = f(x), y_2 = \varphi(x), \dots, y_n = \psi(x)$ be given by a system

$$F(x, y_1, y_2, \dots, y_n) = 0, \quad \Phi(x, y_1, y_2, \dots, y_n) = 0, \quad \dots, \Psi(x, y_1, y_2, \dots, y_n) = 0 \tag{6.56a}$$

of n equations. Differentiation of (6.56a) using (6.51b) results in

$$\left.\begin{array}{l} F_x + F_{y_1}y_1' + F_{y_2}y_2' + \cdots + F_{y_n}y_n' = 0, \\ \Phi_x + \Phi_{y_1}y_1' + \Phi_{y_2}y_2' + \cdots + \Phi_{y_n}y_n' = 0, \\ \vdots + \vdots + \vdots + \vdots + \vdots = 0, \\ \Psi_x + \Psi_{y_1}y_1' + \Psi_{y_2}y_2' + \cdots + \Psi_{y_n}y_n' = 0. \end{array}\right\} \tag{6.56b}$$

Solving (6.56b) yields the derivatives $y_1', y_2', \dots, y_n'$, which are to be looking for. In the same way one can calculate the higher-order derivatives.

5. Two Functions $u = f(x, y), v = \varphi(x, y)$ of Two Variables are given by the system of equations

$$F(x, y, u, v) = 0 \quad \text{and} \quad \Phi(x, y, u, v) = 0. \tag{6.57a}$$

Then differentiation of (6.57a) with respect to x and y with the help of (6.52b) results in

$$\left.\begin{aligned}\frac{\partial F}{\partial x} + \frac{\partial F}{\partial u}\frac{\partial u}{\partial x} + \frac{\partial F}{\partial v}\frac{\partial v}{\partial x} = 0,\\ \frac{\partial \Phi}{\partial x} + \frac{\partial \Phi}{\partial u}\frac{\partial u}{\partial x} + \frac{\partial \Phi}{\partial v}\frac{\partial v}{\partial x} = 0,\end{aligned}\right\} \tag{6.57b}$$

$$\left.\begin{aligned}\frac{\partial F}{\partial y} + \frac{\partial F}{\partial u}\frac{\partial u}{\partial y} + \frac{\partial F}{\partial v}\frac{\partial v}{\partial y} = 0,\\ \frac{\partial \Phi}{\partial y} + \frac{\partial \Phi}{\partial u}\frac{\partial u}{\partial y} + \frac{\partial \Phi}{\partial v}\frac{\partial v}{\partial y} = 0.\end{aligned}\right\} \tag{6.57c}$$

Solving the system (6.57b) for $\frac{\partial u}{\partial x}, \frac{\partial v}{\partial x}$ and the system (6.57c) for $\frac{\partial u}{\partial y}, \frac{\partial v}{\partial y}$ give the first-order partial derivatives. The higher-order derivatives should be calculated in the same way.

6. n Functions of m Variables Given by a System of n Equations The first-order and higher-order partial derivatives can be calculated in the same way as in the previous cases.

6.2.4 Substitution of Variables in Differential Expressions and Coordinate Transformations

6.2.4.1 Function of One Variable

Suppose, given a function $y(x)$ and a differential expression F containing the independent variable, the function, and its derivatives:

$$y = f(x), \tag{6.58a}$$

$$F = F\left(x, y, \frac{dy}{dx}, \frac{d^2y}{dx^2}, \frac{d^3y}{dx^3}, \ldots\right). \tag{6.58b}$$

If the variables are substituted, then the derivatives can be calculated in the following way:

Case 1a: The variable x is replaced by the variable t, and they have the relation

$$x = \varphi(t). \tag{6.59a}$$

Then holds

$$\frac{dy}{dx} = \frac{1}{\varphi'(t)}\frac{dy}{dt}, \quad \frac{d^2y}{dx^2} = \frac{1}{[\varphi'(t)]^3}\left\{\varphi'(t)\frac{d^2y}{dt^2} - \varphi''(t)\frac{dy}{dt}\right\}, \tag{6.59b}$$

$$\frac{d^3y}{dx^3} = \frac{1}{[\varphi'(t)]^5}\left\{[\varphi'(t)]^2\frac{d^3y}{dt^3} - 3\,\varphi'(t)\,\varphi''(t)\frac{d^2y}{dt^2} + [3[\varphi''(t)]^2 - \varphi'(t)\,\varphi'''(t)]\frac{dy}{dx}\right\}, \ldots. \tag{6.59c}$$

Case 1b: If the relation between the variables is not explicit but it is given in implicit form

$$\Phi(x, t) = 0, \tag{6.60}$$

then the derivatives $\frac{dy}{dx}, \frac{d^2y}{dx^2}, \frac{d^3y}{dx^3}$ are calculated by the same formulas, but the derivatives $\varphi'(t), \varphi''(t)$, $\varphi'''(t)$ must be calculated according to the rules for implicit functions. In this case it can happen that the relation (6.58b) contains the variable x. To eliminate x, the relation (6.60) is used.

Case 2: If the function y is replaced by a function $u(x)$, and the relation between them is

$$y = \varphi(u), \tag{6.61a}$$

then the calculation of the derivatives can be performed using the following formulas:

$$\frac{dy}{dx} = \varphi'(u)\frac{du}{dx}, \quad \frac{d^2y}{dx^2} = \varphi'(u)\frac{d^2u}{dx^2} + \varphi''(u)\left(\frac{du}{dx}\right)^2, \tag{6.61b}$$

$$\frac{d^3y}{dx^3} = \varphi'(u)\frac{d^3u}{dx^3} + 3\varphi''(u)\frac{du}{dx}\frac{d^2u}{dx^2} + \varphi'''(u)\left(\frac{du}{dx}\right)^3, \ldots. \tag{6.61c}$$

Case 3: The variables x and y are replaced by the new variables t and u, and the relations between them are given by

$$x = \varphi(t,u), \quad y = \psi(t,u). \tag{6.62a}$$

For the calculation of the derivatives the following formulas are used:

$$\frac{dy}{dx} = \frac{\frac{\partial \psi}{\partial t} + \frac{\partial \psi}{\partial u}\frac{du}{dt}}{\frac{\partial \varphi}{\partial t} + \frac{\partial \varphi}{\partial u}\frac{du}{dt}}, \tag{6.62b}$$

$$\frac{d^2y}{dx^2} = \frac{d}{dx}\left(\frac{dy}{dx}\right) = \frac{d}{dx}\left[\frac{\frac{\partial \psi}{\partial t} + \frac{\partial \psi}{\partial u}\frac{du}{dt}}{\frac{\partial \varphi}{\partial t} + \frac{\partial \varphi}{\partial u}\frac{du}{dt}}\right] = \frac{1}{\frac{\partial \varphi}{\partial t} + \frac{\partial \varphi}{\partial u}\frac{du}{dt}}\frac{d}{dt}\left[\frac{\frac{\partial \psi}{\partial t} + \frac{\partial \psi}{\partial u}\frac{du}{dt}}{\frac{\partial \varphi}{\partial t} + \frac{\partial \varphi}{\partial u}\frac{du}{dt}}\right], \tag{6.62c}$$

$$\frac{1}{B}\frac{d}{dt}\left(\frac{A}{B}\right) = \frac{1}{B^3}\left(B\frac{dA}{dt} - A\frac{dB}{dt}\right), \tag{6.62d}$$

with $A = \frac{\partial \psi}{\partial t} + \frac{\partial \psi}{\partial u}\frac{du}{dt}$ (6.62e) and $B = \frac{\partial \varphi}{\partial t} + \frac{\partial \varphi}{\partial u}\frac{du}{dt}$. (6.62f)

The determination of the third derivative $\frac{d^3y}{dx^3}$ can be done in an analogous way.

■ For the transformation from Cartesian coordinates into polar coordinates according to

$$x = \rho\cos\varphi, \quad y = \rho\sin\varphi \tag{6.63a}$$

the first and second derivatives should be calculated as follows:

$$\frac{dy}{dx} = \frac{\rho'\sin\varphi + \rho\cos\varphi}{\rho'\cos\varphi - \rho\sin\varphi}, \quad (6.63b) \qquad \frac{d^2y}{dx^2} = \frac{\rho^2 + 2\rho'^2 - \rho\rho''}{(\rho'\cos\varphi - \rho\sin\varphi)^3}. \quad (6.63c)$$

6.2.4.2 Function of Two Variables

Suppose given a function $\omega(x,y)$ and a differential expression F containing the independent variables, the function and its partial derivatives:

$$\omega = f(x,y), \quad (6.64a) \qquad F = F\left(x, y, \omega, \frac{\partial \omega}{\partial x}, \frac{\partial \omega}{\partial y}, \frac{\partial^2 \omega}{\partial x^2}, \frac{\partial^2 \omega}{\partial x \partial y}, \frac{\partial^2 \omega}{\partial y^2}, \dots\right). \quad (6.64b)$$

If x and y are replaced by the new variables u and v given by the relations

$$x = \varphi(u,v), \quad y = \psi(u,v), \tag{6.65a}$$

then the first-order partial derivatives can be expressed from the system of equations

$$\frac{\partial \omega}{\partial u} = \frac{\partial \omega}{\partial x}\frac{\partial \varphi}{\partial u} + \frac{\partial \omega}{\partial y}\frac{\partial \psi}{\partial u}, \quad \frac{\partial \omega}{\partial v} = \frac{\partial \omega}{\partial x}\frac{\partial \varphi}{\partial v} + \frac{\partial \omega}{\partial y}\frac{\partial \psi}{\partial v} \tag{6.65b}$$

with the new functions A, B, C, and D of the new variables u and v

$$\frac{\partial \omega}{\partial x} = A\frac{\partial \omega}{\partial u} + B\frac{\partial \omega}{\partial v}, \quad \frac{\partial \omega}{\partial y} = C\frac{\partial \omega}{\partial u} + D\frac{\partial \omega}{\partial v}. \tag{6.65c}$$

The second-order partial derivatives are calculated with the same formulas, only without using ω in them but its partial derivatives $\frac{\partial \omega}{\partial x}$ and $\frac{\partial \omega}{\partial y}$, e.g.,

$$\frac{\partial^2 \omega}{\partial x^2} = \frac{\partial}{\partial x}\left(\frac{\partial \omega}{\partial x}\right) = \frac{\partial}{\partial x}\left(A\frac{\partial \omega}{\partial u} + B\frac{\partial \omega}{\partial v}\right) = A\left(A\frac{\partial^2 \omega}{\partial u^2} + B\frac{\partial^2 \omega}{\partial u \partial v} + \frac{\partial A}{\partial u}\frac{\partial \omega}{\partial u} + \frac{\partial B}{\partial u}\frac{\partial \omega}{\partial v}\right)$$

$$+B\left(A\frac{\partial^2\omega}{\partial u\partial v}+B\frac{\partial^2\omega}{\partial v^2}+\frac{\partial A}{\partial v}\frac{\partial\omega}{\partial u}+\frac{\partial B}{\partial v}\frac{\partial\omega}{\partial v}\right). \quad (6.66)$$

The higher partial derivatives can be calculated in the same way.

■ The Laplace operator (see 13.2.6.5, p. 716) is to be expressed in polar coordinates (see 3.5.2.1, **2.**, p. 191):

$$\Delta\omega=\frac{\partial^2\omega}{\partial x^2}+\frac{\partial^2\omega}{\partial y^2}, \quad (6.67a) \qquad x=\rho\cos\varphi, \quad y=\rho\sin\varphi. \quad (6.67b)$$

The calculations are:

$$\frac{\partial\omega}{\partial\rho}=\frac{\partial\omega}{\partial x}\cos\varphi+\frac{\partial\omega}{\partial y}\sin\varphi, \qquad \frac{\partial\omega}{\partial\varphi}=-\frac{\partial\omega}{\partial x}\rho\sin\varphi+\frac{\partial\omega}{\partial y}\rho\cos\varphi,$$

$$\frac{\partial\omega}{\partial x}=\cos\varphi\frac{\partial\omega}{\partial\rho}-\frac{\sin\varphi}{\rho}\frac{\partial\omega}{\partial\varphi}, \qquad \frac{\partial\omega}{\partial y}=\sin\varphi\frac{\partial\omega}{\partial\rho}+\frac{\cos\varphi}{\rho}\frac{\partial\omega}{\partial\varphi},$$

$$\frac{\partial^2\omega}{\partial x^2}=\cos\varphi\frac{\partial}{\partial\rho}\left(\cos\varphi\frac{\partial\omega}{\partial\rho}-\frac{\sin\varphi}{\rho}\frac{\partial\omega}{\partial\varphi}\right)-\frac{\sin\varphi}{\rho}\frac{\partial}{\partial\varphi}\left(\cos\varphi\frac{\partial\omega}{\partial\rho}-\frac{\sin\varphi}{\rho}\frac{\partial\omega}{\partial\varphi}\right).$$

Similarly, $\dfrac{\partial^2\omega}{\partial y^2}$ is calculated, so finally:

$$\Delta\omega=\frac{\partial^2\omega}{\partial\rho^2}+\frac{1}{\rho^2}\frac{\partial^2\omega}{\partial\varphi^2}+\frac{1}{\rho}\frac{\partial\omega}{\partial\rho}. \quad (6.67c)$$

Remark: If functions of more than two variables should be substituted, then similar substitution formulas can be derived.

6.2.5 Extreme Values of Functions of Several Variables

6.2.5.1 Definition of a Relative Extreme Value

A function $u=f(x_1,x_2,\dots,x_i,\dots,x_n)$ has a relative extreme value at a point $P_0(x_{10},x_{20},\dots,x_{i0},\dots,x_{n0})$, if there is a number ϵ such that for every point $P(x_1,x_2,\dots,x_n)$ belonging to the domain $x_{10}-\epsilon<x_1<x_{10}+\epsilon, x_{20}-\epsilon<x_2<x_{20}+\epsilon,\dots,x_{n0}-\epsilon<x_n<x_{n0}+\epsilon$ and to the domain of the function but different from P_0, then for a maximum the inequality

$$f(x_1,x_2,\dots,x_n)<f(x_{10},x_{20},\dots,x_{n0}) \quad (6.68a)$$

holds, and for a minimum the inequality

$$f(x_1,x_2,\dots,x_n)>f(x_{10},x_{20},\dots,x_{n0}) \quad (6.68b)$$

holds. Using the terminology of several dimensional spaces (see 2.18.1, p. 118) a function has a relative maximum or a relative minimum at a point if it is greater or smaller there than at the neighboring points.

6.2.5.2 Geometric Representation

In the case of a function of two variables, represented in a Cartesian coordinate system as a surface (see 2.18.1.2, p. 119), the relative extreme value geometrically means that the applicate (see 3.5.3.1, **3.**, p. 210) of the surface in the point A is greater or smaller than the applicate of the surface in any other point in a sufficiently small neighborhood of A **(Fig. 6.16)**.

If the surface has a relative extremum at the point P_0 which is an interior point of its domain, and if the surface has a tangent plane at this point, then the tangent plane is parallel to the x,y plane **(Fig. 6.16a,b)**. This property is necessary but not sufficient for a maximum or minimum at a point

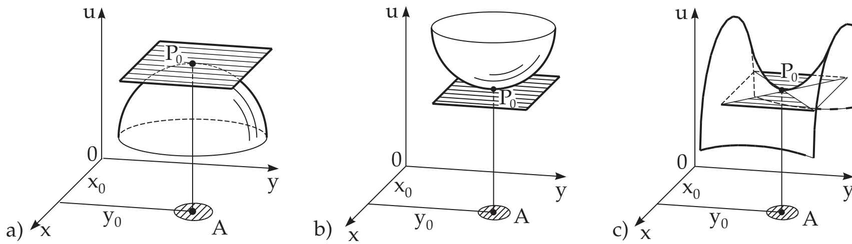


Figure 6.16

P_0. For example **Fig. 6.16c** shows a surface having a horizontal tangent plane at P_0, but there is a saddle point here and not an extremum.

6.2.5.3 Determination of Extreme Values of Differentiable Functions of Two Variables

If $u = f(x, y)$ is given, then one solves the system of equations $f_x = 0$, $f_y = 0$. The resulting pairs of values $(x_1, y_1), (x_2, y_2), \ldots$ can be substituted into the second derivatives

$$A = \frac{\partial^2 f}{\partial x^2}, \quad B = \frac{\partial^2 f}{\partial x \partial y}, \quad C = \frac{\partial^2 f}{\partial y^2}. \tag{6.69}$$

Depending on the expression

$$\Delta = \begin{vmatrix} A & B \\ B & C \end{vmatrix} = AC - B^2 = [f_{xx}f_{yy} - (f_{xy})^2]_{x=x_i, y=y_i} \quad (i = 1, 2, \ldots) \tag{6.70}$$

it can be decided whether an extreme value exists and of what kind it is:

1. In the case $\Delta > 0$ the function $f(x, y)$ has an extreme value at (x_i, y_i), and for $f_{xx} < 0$ it is a maximum, for $f_{xx} > 0$ it is a minimum (sufficient condition).

2. In the case $\Delta < 0$ the function $f(x, y)$ does not have an extremum.

3. In the case $\Delta = 0$, one needs further investigation.

6.2.5.4 Determination of the Extreme Values of a Function of n Variables

If $u = f(x_1, x_2, \ldots, x_n)$ is given, then first it is to find a solution $(x_{10}, x_{20}, \ldots, x_{n0})$ of the system of the n equations

$$f_{x_1} = 0, \ f_{x_2} = 0, \ \ldots, f_{x_n} = 0, \tag{6.71}$$

because it is a necessary condition for an extreme value. (6.71) is not a sufficient condition. Therefore one prepares a matrix of the second-order partial derivatives such that $a_{ij} = \dfrac{\partial^2 f}{\partial x_i \partial x_j}$. Then it is to substitute a solution of the system of equations (6.71) into the terms, and to prepare the sequence of left upper subdeterminants $(a_{11},\ a_{11}a_{22} - a_{12}a_{21},\ \ldots)$. Then there are the following cases:

1. The signs of the subdeterminants follow the rule $-, +, -, +, \ldots$, then there is a maximum there.

2. The signs of the subdeterminants follow the rule $+, +, +, +, \ldots$, then there is a minimum there.

3. There are some zero values among the subdeterminants, but the signs of the non-zero subdeterminants coincide with the signs of the corresponding positions of one of the first two cases. Then further investigation is required: Usually one checks the values of the function in a close neighborhood of $x_{10}, x_{20}, \ldots, x_{n0}$.

4. The signs of the subdeterminants does not follow the rules given in cases 1. and 2.: There is no extremum at that point.

The case of two variables is of course a special case of the case of n variables, (see [6.4]).

6.2.5.5 Solution of Approximation Problems

Several different approximation problems can be solved with the help of the determination of the extreme values of functions of several variables, e.g., *fitting problems* or *mean squares problems.*

Problems to solve:

- Determination of Fourier coefficients (see 7.4.1.2, p. 475, 19.6.4.1, p. 992).
- Determination of the coefficients and parameters of approximation functions (see 19.6.2, p. 984 ff).
- Determination of an approximate solution of an overdetermined linear system of equations (see 19.2.1.3, p. 958).

Methods: For these problems the following methods are used:

- Gaussian least squares method (see, e.g., 19.6.2, p. 984).
- Least squares method (see 19.6.2.2, p. 985).
- Approximation in mean square (continuous and discrete) (see, e.g., 19.6.2, p.984).
- Calculus of observations (or fitting) (see 19.6.2, p. 984) and regression (see 16.3.4.2, **1.**, p. 841).

6.2.5.6 Extreme Value Problem with Side Conditions

To determine are the extreme values of a function $u = f(x_1, x_2, \ldots, x_n)$ of n variables with the side conditions

$$\varphi(x_1, x_2, \ldots, x_n) = 0,\ \psi(x_1, x_2, \ldots, x_n) = 0, \ldots,\ \chi(x_1, x_2, \ldots, x_n) = 0. \tag{6.72a}$$

Because of the conditions, the variables are not independent, and if the number of conditions is k, obviously $k < n$ must hold. One possibility to determine the extreme values of u is to express k variables with the others from the system of equations of the conditions, to substitute them into the original function. Then the result is an extreme value problem without conditions for $n-k$ variables. The other way is the *Lagrange multiplier method.* Introducing k undefined multipliers $\lambda, \mu, \ldots, \kappa$, and composing the *Lagrange function (Lagrangian)* of $n + k$ variables $x_1, x_2, \ldots, x_n, \lambda, \mu, \ldots, \kappa$ gives:

$$\begin{aligned}\Phi\ (x_1, x_2, \ldots, x_n, \lambda, \mu, \ldots, \kappa) \\ = f(x_1, x_2, \ldots, x_n) + \lambda\, \varphi(x_1, x_2, \ldots, x_n) + \mu\, \psi(x_1, x_2, \ldots, x_n) + \cdots \\ + \kappa\, \chi(x_1, x_2, \ldots, x_n).\end{aligned} \tag{6.72b}$$

The necessary condition for an extremum of the function Φ is a system of $n + k$ equations (6.71) with the unknowns $x_1, x_2, \ldots, x_n, \lambda, \mu, \ldots, \kappa$:

$$\varphi = 0, \psi = 0,\ \ldots,\ \chi = 0,\ \Phi_{x_1} = 0,\ \Phi_{x_2} = 0, \ldots, \Phi_{x_n} = 0 \tag{6.72c}$$

The necessary condition for an extremum of the function f at the point $P_0(x_{10}, x_{20}, \ldots, x_{n0})$ with the side conditions (6.72a) is that the system of values $x_{10}, x_{20}, \ldots, x_{n0}$ must fulfill the equations (6.72c). So it is to look for the extremum points of f among the solutions $x_{10}, x_{20}, \ldots, x_{n0}$ of the system of equations (6.72c). To determine whether there are really extreme values at these points fulfilling the necessary conditions requires further investigations, for which the general rules are fairly complicated. Usually one uses some appropriate and individual calculations depending on the function f to prove if there is an extremum, or not. Often approximation calculations are helpful, i.e., the comparison with values of the function in the neighborhood of P_0.

■ The extreme value of the function $u = f(x, y)$ with the side condition $\varphi(x, y) = 0$ will be determined from the three equations

$$\varphi(x, y) = 0,\ \frac{\partial}{\partial x}[f(x, y) + \lambda\varphi(x, y)] = 0,\ \frac{\partial}{\partial y}[f(x, y) + \lambda\varphi(x, y)] = 0. \tag{6.73}$$

There are three unknowns, x, y, λ. Since the three equations (6.73) are only necessary but not sufficient conditions for the existence of an extremum, further investigation is needed whether there is an extremum at the solution of this system. A mathematical criterion is rather complicated (see [6.3], [6.8]); comparisons of the values of the function at points in the close neighborhood are often useful.

7 Infinite Series

7.1 Sequences of Numbers

7.1.1 Properties of Sequences of Numbers

7.1.1.1 Definition of Sequence of Numbers

An *infinite sequence of numbers* (hereafter briefly sequence of numbers) is an infinite system of numbers

$$a_1, a_2, \ldots, a_n, \ldots \quad \text{or briefly} \ \{a_n\} \ \text{with} \ n = 1, 2, \ldots \, , \tag{7.1}$$

arranged in a given order. The numbers of the sequence of numbers are called the *terms of the sequence*. Among the *terms of a sequence* of numbers the same numbers can occur several times. A sequence is considered to be defined if *the law of formation*, i.e., a rule is given, by which any term of the sequence can be uniquely determined. Mostly there is a formula for the general term a_n.

Examples of Sequences of Numbers:

■ **A:** $a_n = n$: $1, 2, 3, 4, 5, \ldots$.

■ **B:** $a_n = 4 + 3(n-1)$: $4, 7, 10, 13, 16, \ldots$.

■ **C:** $a_n = 3\left(-\frac{1}{2}\right)^{n-1}$: $3, -\frac{3}{2}, \frac{3}{4}, -\frac{3}{8}, \frac{3}{16}, \ldots$.

■ **D:** $a_n = (-1)^{n+1}$: $1, -1, 1, -1, 1, \ldots$.

■ **E:** $a_n = 3 - \frac{1}{2^{n-2}}$: $1, 2, 2\frac{1}{2}, 2\frac{3}{4}, 2\frac{7}{8}, \ldots$ (read $2\frac{3}{4} = \frac{11}{4}$).

■ **F:** $a_n = 3\frac{1}{3} - \frac{1}{3} \cdot 10^{-\frac{n-1}{2}}$ for odd n and

$a_n = 3\frac{1}{3} + \frac{2}{3} \cdot 10^{-\frac{n}{2} + 1}$ for even n: $3; 4; 3.3; 3.4; 3.33; 3.34; 3.333; 3.334; \ldots$.

■ **G:** $a_n = \frac{1}{n}$: $1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \ldots$.

■ **H:** $a_n = (-1)^{n+1} n$: $1, -2, 3, -4, 5, -6, \ldots$.

■ **I:** $a_n = -\frac{n+1}{2}$ for odd n and $a_n = 0$ for even n: $-1, 0, -2, 0, -3, 0, -4, 0, \ldots$.

■ **J:** $a_n = 3 - \frac{1}{2^{\frac{n}{2} - \frac{3}{2}}}$ for odd n and $a_n = 13 - \frac{1}{2^{\frac{n}{2}} - 2}$ for even n: $1, 11, 2, 12, 2\frac{1}{2}, 12\frac{1}{2}, 2\frac{3}{4}, 12\frac{3}{4}, \ldots$.

7.1.1.2 Monotone Sequences of Numbers

A sequence of numbers $a_1, a_2, \ldots, a_n, \ldots$ is *monotone increasing* if

$$a_1 \le a_2 \le a_3 \le \cdots \le a_n \le \cdots \, , \tag{7.2}$$

is valid and it is *monotone decreasing* if

$$a_1 \ge a_2 \ge a_3 \ge \cdots \ge a_n \ge \cdots \tag{7.3}$$

is valid. One talks about a *strictly monotone increasing sequence of numbers* or *strictly monotone decreasing sequence of numbers*, if equality never holds in (7.2) or (7.3).

Examples of Monotone Sequences of Numbers:

■ **A:** Among the sequences of numbers from **A** to **J** the sequences **A, B, E** are strictly monotone increasing.

■ **B:** The sequence of numbers **G** is strictly monotone decreasing.

7.1.1.3 Bounded Sequences of Numbers

A sequence of numbers is called *bounded* if for all terms

$$|a_n| < K \tag{7.4}$$

I.N. Bronshtein et al., *Handbook of Mathematics*,
DOI 10.1007/978-3-662-46221-8_7

is valid for a certain $K > 0$. If such a K (*bound*) does not exist, then the sequence of numbers is *unbounded*.

■ Among the sequences of numbers from **A** to **J** the sequences of numbers **C** with $K = 4$, **D** with $K = 2$, **E** with $K = 3$, **F** with $K = 5$, **G** with $K = 2$ and **J** with $K = 13$ are bounded.

7.1.2 Limits of Sequences of Numbers

1. Limit of a Sequence of Numbers

An infinite sequence of numbers (7.1) has a *limit* A if for an unlimited increase of the index n the difference $a_n - A$ becomes arbitrarily small. Precisely defined this means: For an arbitrarily small $\varepsilon > 0$ there exists an index $n_0(\varepsilon)$ such that for every $n > n_0$

$$|a_n - A| < \varepsilon. \tag{7.5a}$$

The sequence of numbers has the limit $+\infty$ $(-\infty)$, if for arbitrary $K > 0$ there exists an index $n_0(K)$ such that for every $n > n_0$

$$a_n > K \qquad (a_n < -K). \tag{7.5b}$$

2. Convergence of a Sequence of Numbers

If a sequence of numbers $\{a_n\}$ satisfies (7.5a), then one says it *converges to* A. This is denoted by

$$\lim_{n\to\infty} a_n = A \quad \text{or} \quad a_n \to A. \tag{7.6}$$

■ Among the sequences of numbers from **A** to **J** on the previous page, **C** with $A = 0$, **E** with $A = 3$, **F** with $A = 3\frac{1}{3}$, **G** with $A = 0$ are convergent.

3. Divergence of a Sequence of Numbers

Non-convergent sequences of numbers are called *divergent*. One talks about *proper divergence* in the case of (7.5b), i.e., if as n exceeds any value, a_n exceeds any large positive number K $(K > 0)$ so that it never goes below, or if as n exceeds any value, a_n goes below any negative number $-K$ $(K > 0)$ with arbitrarily large magnitude and never increases above it, i.e., if it has the limit $\pm\infty$. Then one writes:

$$\lim_{n\to\infty} a_n = \infty \quad (a_n > K\,\forall n > n_0) \quad \text{or} \quad \lim_{n\to-\infty} a_n = -\infty \quad (a_n < -K\,\forall n > n_0). \tag{7.7}$$

Otherwise the sequence of numbers is called *improperly divergent*.

Examples of Divergent Sequences of Numbers:

■ **A:** Among the sequences of numbers from **A** to **J** on the previous page, **A** and **B** tend $+\infty$, they are properly divergent.

■ **B:** Among the sequences of numbers **D** is improperly divergent.

4. Theorems for Limits of Sequences of Numbers

a) If the sequences of numbers $\{a_n\}$ and $\{b_n\}$ are convergent, then

$$\lim_{n\to\infty}(a_n + b_n) = \lim_{n\to\infty} a_n + \lim_{n\to\infty} b_n, \tag{7.8}$$

$$\lim_{n\to\infty}(a_n b_n) = (\lim_{n\to\infty} a_n)(\lim_{n\to\infty} b_n) \tag{7.9}$$

hold, and if $b_n \neq 0$ for every n, and $\lim\limits_{n\to\infty} b_n \neq 0$, then holds

$$\lim_{n\to\infty} \frac{a_n}{b_n} = \frac{\lim\limits_{n\to\infty} a_n}{\lim\limits_{n\to\infty} b_n}. \tag{7.10}$$

Remark: If $\lim\limits_{n\to\infty} b_n = 0$ and $\{a_n\}$ is bounded, then $\lim\limits_{n\to\infty}(a_n b_n) = 0$ even if $\{a_n\}$ does not have any finite limit.

b) If $\lim\limits_{n\to\infty} a_n = \lim\limits_{n\to\infty} b_n = A$ is valid and at least, from an index n_1 and beyond, the inequality $a_n \leq c_n \leq b_n$ holds, then also holds

$$\lim_{n\to\infty} c_n = A. \tag{7.11}$$

c) A monotone and bounded sequence of numbers has a finite limit. If a monotone increasing sequence of numbers $a_1 \leq a_2 \leq a_3 \leq \ldots$ is bounded above, i.e., $a_n \leq K_1$ for all n, then it is convergent, and its limit is equal to its *least upper bound* which is the smallest possible value for K_1. If a monotone decreasing sequence of numbers $a_1 \geq a_2 \geq a_3 \geq \ldots$ is bounded below, i.e., $a_n \geq K_2$ for all n, then it is convergent, and its limit is equal to its *greatest lower bound* which is the largest possible value for K_2.

7.2 Number Series

7.2.1 General Convergence Theorems

7.2.1.1 Convergence and Divergence of Infinite Series

1. Infinite Series and its Sum

From the terms a_k of an *infinite sequence of numbers* $\{a_k\}$ (see 7.1.1.1, p. 457) the formal expression

$$a_1 + a_2 + \cdots + a_n + \cdots = \sum_{k=1}^{\infty} a_k \tag{7.12}$$

can be composed and this is called an *infinite series* (hereafter briefly series); a_k is the *general term* of the series. The finite sums

$$S_1 = a_1,\ \ S_2 = a_1 + a_2,\ \ \ldots,\ \ S_n = \sum_{k=1}^{n} a_k \tag{7.13}$$

are called *partial sums*.

2. Convergent and Divergent Series

A series (7.12) is called *convergent* if the sequence of partial sums $\{S_n\}$ is convergent. The *limit*

$$S = \lim_{n\to\infty} S_n = \sum_{k=1}^{\infty} a_k \tag{7.14}$$

is called the *sum* of the series. If the limit (7.14) does not exist or it is equal to $\pm\infty$, then the series is called *divergent*. In this case the partial sums are not bounded or they oscillate. So the determination of the convergence of an infinite series is reduced to the determination of the limit of a sequence $\{S_n\}$.

■ **A:** The *geometric series* (see 1.2.3, p. 19)

$$1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \cdots + \frac{1}{2^n} + \cdots \tag{7.15}$$

is convergent with the sum $S = 2$ (see (1.54b), p. 19) with $a_0 = 1, q = 1/2$).

■ **B:** The *harmonic series* (7.16) and the series (7.17) and (7.18)

$$1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n} + \cdots \tag{7.16}$$

$$1 + 1 + 1 + \cdots + 1 + \cdots \tag{7.17}$$

$$1 - 1 + 1 - \cdots + (-1)^{n-1} + \cdots \tag{7.18}$$

are divergent. For the series (7.16) and (7.17) $\lim\limits_{n\to\infty} S_n = \infty$ holds, (7.18) oscillates.

3. Remainder

The *remainder* of a convergent series $S = \sum\limits_{k=1}^{\infty} a_k$ is the difference between its sum S and the partial sum S_n:

$$R_n = S - S_n = \sum_{k=n+1}^{\infty} a_k = a_{n+1} + a_{n+2} + \cdots\,. \tag{7.19}$$

7.2.1.2 General Theorems about the Convergence of Series

1. Necessary Criterion for the Convergence of a Series The sequence of terms of a convergent series is a null sequence:

$$\lim_{n\to\infty} a_n = 0. \qquad (7.20)$$

This is only a *necessary* but not *sufficient condition.*

■ For the harmonic series (7.16) $\lim\limits_{n\to\infty} a_n = 0$ holds, but $\lim\limits_{n\to\infty} S_n = \infty$.

2. Leaving out the Initial Terms If one leaves out finitely many initial terms of a series or one introduces finitely many further terms into it at the begin or if changing the order of finitely many terms, then its convergence behavior does not change. Exchange of the order of finitely many terms does not affect the sum if it exists.

3. Multiplication of all Terms If all terms of a convergent series are multiplied by the same factor c, then the convergence of the series does not change; its sum is multiplied by the factor c.

4. Termwise Addition or Subtraction By adding or subtracting two convergent series

$$a_1 + a_2 + \cdots + a_n + \cdots = \sum_{k=1}^{\infty} a_k = S_1, \ (7.21a) \qquad b_1 + b_2 + \cdots + b_n + \cdots = \sum_{k=1}^{\infty} b_k = S_2 \quad (7.21b)$$

term by term, then the result is a convergent series, and its sum is

$$(a_1 \pm b_1) + (a_2 \pm b_2) + \cdots + (a_n \pm b_n) + \cdots = S_1 \pm S_2. \qquad (7.21c)$$

7.2.2 Convergence Criteria for Series with Positive Terms

7.2.2.1 Comparison Criterion

Suppose there are two series

$$a_1 + a_2 + \cdots + a_n + \cdots = \sum_{n=1}^{\infty} a_n, \qquad (7.22a) \qquad b_1 + b_2 + \cdots + b_n + \cdots = \sum_{n=1}^{\infty} b_n \qquad (7.22b)$$

with only positive terms ($a_n > 0$, $b_n > 0$). If $a_n \geq b_n$ holds from a certain n_0, then the convergence of the series (7.22a) yields the convergence of the series (7.22b), and the divergence of the series (7.22b) yields the divergence of the series (7.22a). In the first case (7.22a) is called a convergent majorant and in the second case (7.22b) a divergent minorant .

■ **A:** Comparing the terms of the series

$$1 + \frac{1}{2^2} + \frac{1}{3^3} + \cdots + \frac{1}{n^n} + \cdots \qquad (7.23a)$$

with the geometric series (7.15), the convergence of the series (7.23a) follows. From $n = 2$ the terms of the series (7.23a) are smaller than the terms of the convergent series (7.15):

$$\frac{1}{n^n} < \frac{1}{2^{n-1}} \quad (n \geq 2). \qquad (7.23b)$$

■ **B:** From the comparison of the terms of the series

$$1 + \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{3}} + \cdots + \frac{1}{\sqrt{n}} + \cdots \qquad (7.24a)$$

with the terms of the harmonic series (7.16) follows the divergence of the series (7.24a). For $n > 1$ the terms of the series (7.24a) are greater than those of the divergent series (7.16):

$$\frac{1}{\sqrt{n}} > \frac{1}{n} \quad (n > 1). \qquad (7.24b)$$

7.2.2.2 D'Alembert's Ratio Test

If for the series

$$a_1 + a_2 + \cdots + a_n + \cdots = \sum_{n=1}^{\infty} a_n \tag{7.25a}$$

all the ratios $\frac{a_{n+1}}{a_n}$ are smaller than a number $q < 1$ from a certain n_0 onwards, then the series is convergent:

$$\frac{a_{n+1}}{a_n} \le q < 1 \quad (n \ge n_0)\,. \tag{7.25b}$$

If these ratios are greater than a number $Q > 1$ from a certain n_0 onwards, then the series is divergent. From the two previous statements it follows that if the limit

$$\lim_{n\to\infty} \frac{a_{n+1}}{a_n} = \rho \tag{7.25c}$$

exists, then for $\rho < 1$ the series is convergent, for $\rho > 1$ it is divergent. In the case $\rho = 1$ the ratio test gives no information whether the series is convergent or not.

■ **A**: The series $\frac{1}{2} + \frac{2}{2^2} + \frac{3}{2^3} + \cdots + \frac{n}{2^n} + \cdots$ (7.26a)

is convergent, because

$$\rho = \lim_{n\to\infty} \left(\frac{n+1}{2^{n+1}} : \frac{n}{2^n} \right) = \lim_{n\to\infty} \frac{1 + \frac{1}{n}}{2} = \frac{1}{2} < 1 \text{ holds.} \tag{7.26b}$$

■ **B**: For the series $2 + \frac{3}{4} + \frac{4}{9} + \cdots + \frac{n+1}{n^2} + \cdots$ (7.27a)

the ratio test does not give any result about whether the series is convergent or not, because

$$\rho = \lim_{n\to\infty} \left(\frac{n+2}{(n+1)^2} : \frac{n+1}{n^2} \right) = 1. \tag{7.27b}$$

7.2.2.3 Root Test of Cauchy

If for a series

$$a_1 + a_2 + \cdots + a_n + \cdots = \sum_{n=1}^{\infty} a_n \tag{7.28a}$$

from a certain n_0 onwards for every value $\sqrt[n]{a_n}$

$$\sqrt[n]{a_n} < q < 1 \tag{7.28b}$$

holds, then the series is convergent. If from a certain n_0 every value $\sqrt[n]{a_n}$ is greater than a number Q where $Q > 1$ holds, then the series is divergent.
From the previous statements it follows that if

$$\lim_{n\to\infty} \sqrt[n]{a_n} = \rho \tag{7.28c}$$

exists, in the case $\rho < 1$ the series is convergent, in the case $\rho > 1$ it is divergent. For $\rho = 1$ with this test one cannot tell anything about the convergence behavior of the series.

■ The series $\frac{1}{2} + \left(\frac{2}{3}\right)^4 + \left(\frac{3}{4}\right)^9 + \cdots + \left(\frac{n}{n+1}\right)^{n^2} + \cdots$ (7.29a)

is convergent because

$$\rho = \lim_{n\to\infty} \sqrt[n]{\left(\frac{n}{n+1}\right)^{n^2}} = \lim_{n\to\infty}\left(\frac{1}{1+\frac{1}{n}}\right)^n = \frac{1}{e} < 1 \quad \text{holds.} \tag{7.29b}$$

7.2.2.4 Integral Test of Cauchy

1. Convergence If a series has the general term $a_n = f(n)$, and $f(x)$ is a monotone decreasing function such that the improper integral

$$\int_c^\infty f(x)\,dx \qquad \text{(see 8.2.3.2, \textbf{1.}, p. 507)} \tag{7.30}$$

exists (it is convergent), then the series is convergent.

2. Divergence If the above integral (7.30) is divergent, then the series with the general term $a_n = f(n)$ is divergent, too.
The lower limit c of the integral is almost arbitrary but it must be chosen so that for $c < x < \infty$ the function $f(x)$ should be monotone decreasing.

■ The series (7.27a) is divergent because

$$f(x) = \frac{x+1}{x^2}, \quad \int_c^\infty \frac{x+1}{x^2}\,dx = \left[\ln x - \frac{1}{x}\right]_c^\infty = \infty. \tag{7.31}$$

7.2.3 Absolute and Conditional Convergence

7.2.3.1 Definition

Along with the series (7.12) whose terms can have different signs, one considers also the series

$$|a_1| + |a_2| + \cdots + |a_n| + \cdots = \sum_{n=1}^{\infty} |a_n|, \tag{7.32}$$

whose terms are the absolute values of the terms of the original sequence (7.12). If the series (7.32) is convergent, then the original one (7.12) is convergent, too. (This statement is valid also for series with complex terms.) In this case, the series (7.12) is called *absolutely convergent.* If the series (7.32) is divergent, then the series (7.12) can be either divergent or convergent. In the second case, the series (7.12) is called *conditionally convergent.*

■ **A**: The series $\dfrac{\sin\alpha}{2} + \dfrac{\sin 2\alpha}{2^2} + \cdots + \dfrac{\sin n\alpha}{2^n} + \cdots,$ (7.33a)

where α is an arbitrarily constant number, is absolutely convergent, because the series of absolute values with the general term $\left|\dfrac{\sin n\alpha}{2^n}\right|$ is convergent. This is obvious by comparing it with the geometric series (7.15):

$$\left|\frac{\sin n\alpha}{2^n}\right| \le \frac{1}{2^n}. \tag{7.33b}$$

■ **B**: The series $1 - \dfrac{1}{2} + \dfrac{1}{3} - \cdots + (-1)^{n-1}\dfrac{1}{n} + \cdots$ (7.34)

is conditionally convergent, because it is convergent according to (7.36b), and the series made of the absolute values of the terms is the divergent harmonic series (7.16) whose general term is $\frac{1}{n} = |a_n|$.

7.2.3.2 Properties of Absolutely Convergent Series

1. Exchange of Terms

a) The terms of an absolutely convergent series can be exchanged with each other arbitrarily (even infinitely many of them) and the sum does not change.

b) Exchanging an infinite number of terms of a conditionally convergent series can change the sum and even the convergence behavior. *Theorem of Riemann*: The terms of a conditionally convergent series can be rearranged so that the sum will be equal to any given value, even to $\pm\infty$.

2. Addition and Subtraction

Absolutely convergent series can be added and subtracted term-by-term; the result is absolutely convergent.

3. Multiplication

Multiplying a sum by a sum, the result is a sum of the products where every term of the first factor is multiplied by every term of the second one. These two-term products can be arranged in different ways. The most common way for this arrangement is made as if the series were power series, i.e.:

$$(a_1 + a_2 + \cdots + a_n + \cdots)(b_1 + b_2 + \cdots + b_n + \cdots) \\ = \underbrace{a_1b_1} + \underbrace{a_2b_1 + a_1b_2} + \underbrace{a_3b_1 + a_2b_2 + a_1b_3} + \cdots + \underbrace{a_nb_1 + a_{n-1}b_2 + \cdots + a_1b_n} + \cdots . \quad (7.35a)$$

If two series are absolutely convergent, then their product is absolutely convergent, so it has the same sum in any arrangement. If $\sum a_n = S_a$ and $\sum b_n = S_b$ hold, then the sum of the product is

$$S = S_aS_b. \quad (7.35b)$$

If two series $a_1 + a_2 + \cdots + a_n + \cdots = \sum_{n=1}^{\infty} a_n$ and $b_1 + b_2 + \cdots + b_n + \cdots = \sum_{n=1}^{\infty} b_n$ are convergent, and at least one of them is absolutely convergent, then their product is also convergent, but not necessarily absolutely convergent.

7.2.3.3 Alternating Series

1. Leibniz Alternating Series Test (Theorem of Leibniz)

For an alternating series

$$a_1 - a_2 + a_3 - \cdots \pm a_n \mp \cdots , \quad (7.36a)$$

where a_n are positive numbers, a sufficient condition of convergence is if the following two relations hold:

$$\mathbf{1}.\ \lim_{n\to\infty} a_n = 0 \quad \text{and} \quad \mathbf{2}.\ a_1 > a_2 > a_3 > \cdots > a_n > \cdots . \quad (7.36b)$$

■ The series (7.34) is convergent because of this criterion.

2. Estimation of the Remainder of an Alternating Series

Considering the first n terms of an alternating series, then the remainder R_n has the same sign as the first omitted term a_{n+1}, and the absolute value of R_n is smaller than $|a_{n+1}|$:

$$\operatorname{sign} R_n = \operatorname{sign}(a_{n+1}) \quad \text{with} \quad R_n = S - S_n, \quad (7.37a) \qquad |S - S_n| < |a_{n+1}|. \quad (7.37b)$$

■ Considering the series

$$1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \cdots \pm \frac{1}{n} \mp \cdots = \ln 2 \quad (7.38a) \quad \text{the remainder is } |\ln 2 - S_n| < \frac{1}{n+1}. \quad (7.38b)$$

7.2.4 Some Special Series

7.2.4.1 The Values of Some Important Number Series

$$1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \cdots + \frac{1}{n!} + \cdots = e\,, \tag{7.39}$$

$$1 - \frac{1}{1!} + \frac{1}{2!} - \frac{1}{3!} + \cdots + \frac{(-1)^n}{n!} + \cdots = \frac{1}{e}\,, \tag{7.40}$$

$$1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \cdots + \frac{(-1)^{n+1}}{n} + \cdots = \ln 2\,, \tag{7.41}$$

$$1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \cdots + \frac{1}{2^n} + \cdots = 2\,, \tag{7.42}$$

$$1 - \frac{1}{2} + \frac{1}{4} - \frac{1}{8} + \cdots + \frac{(-1)^n}{2^n} + \cdots = \frac{2}{3}\,, \tag{7.43}$$

$$1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \frac{1}{9} - \cdots + \frac{(-1)^{n-1}}{2n-1} + \cdots = \frac{\pi}{4}\,, \tag{7.44}$$

$$\frac{1}{1\cdot 2} + \frac{1}{2\cdot 3} + \frac{1}{3\cdot 4} + \cdots + \frac{1}{n(n+1)} + \cdots = 1\,, \tag{7.45}$$

$$\frac{1}{1\cdot 3} + \frac{1}{3\cdot 5} + \frac{1}{5\cdot 7} + \cdots + \frac{1}{(2n-1)(2n+1)} + \cdots = \frac{1}{2}\,, \tag{7.46}$$

$$\frac{1}{1\cdot 3} + \frac{1}{2\cdot 4} + \frac{1}{3\cdot 5} + \cdots + \frac{1}{(n-1)(n+1)} + \cdots = \frac{3}{4}\,, \tag{7.47}$$

$$\frac{1}{3\cdot 5} + \frac{1}{7\cdot 9} + \frac{1}{11\cdot 13} + \cdots + \frac{1}{(4n-1)(4n+1)} + \cdots = \frac{1}{2} - \frac{\pi}{8}\,, \tag{7.48}$$

$$\frac{1}{1\cdot 2\cdot 3} + \frac{1}{2\cdot 3\cdot 4} + \cdots + \frac{1}{n(n+1)(n+2)} + \cdots = \frac{1}{4}\,, \tag{7.49}$$

$$\frac{1}{1\cdot 2\ldots l} + \frac{1}{2\cdot 3\ldots(l+1)} + \cdots + \frac{1}{n\ldots(n+l-1)} + \cdots = \frac{1}{(l-1)(l-1)!}\,, \tag{7.50}$$

$$1 + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + \cdots + \frac{1}{n^2} + \cdots = \frac{\pi^2}{6}\,, \tag{7.51}$$

$$1 - \frac{1}{2^2} + \frac{1}{3^2} - \frac{1}{4^2} + \cdots + \frac{(-1)^{n+1}}{n^2} + \cdots = \frac{\pi^2}{12}\,, \tag{7.52}$$

$$\frac{1}{1^2} + \frac{1}{3^2} + \frac{1}{5^2} + \cdots + \frac{1}{(2n+1)^2} + \cdots = \frac{\pi^2}{8}\,, \tag{7.53}$$

$$1 + \frac{1}{2^4} + \frac{1}{3^4} + \frac{1}{4^4} + \cdots + \frac{1}{n^4} + \cdots = \frac{\pi^4}{90}\,, \tag{7.54}$$

$$1 - \frac{1}{2^4} + \frac{1}{3^4} - \cdots + \frac{(-1)^{n+1}}{n^4} + \cdots = \frac{7\pi^4}{720}\,, \tag{7.55}$$

$$\frac{1}{1^4} + \frac{1}{3^4} + \frac{1}{5^4} + \cdots + \frac{1}{(2n+1)^4} + \cdots = \frac{\pi^4}{96}\,, \tag{7.56}$$

$$1 + \frac{1}{2^{2k}} + \frac{1}{3^{2k}} + \frac{1}{4^{2k}} + \cdots + \frac{1}{n^{2k}} + \cdots = \frac{\pi^{2k}2^{2k-1}}{(2k)!}B_k\,,^{*} \tag{7.57}$$

$$1 - \frac{1}{2^{2k}} + \frac{1}{3^{2k}} - \frac{1}{4^{2k}} + \cdots + \frac{(-1)^{n+1}}{n^{2k}} + \cdots = \frac{\pi^{2k}(2^{2k-1}-1)}{(2k)!}B_k\,, \tag{7.58}$$

$$1 + \frac{1}{3^{2k}} + \frac{1}{5^{2k}} + \frac{1}{7^{2k}} + \cdots + \frac{1}{(2n-1)^{2k}} + \cdots = \frac{\pi^{2k}(2^{2k}-1)}{2\cdot(2k)!}B_k\,, \tag{7.59}$$

$$1 - \frac{1}{3^{2k+1}} + \frac{1}{5^{2k+1}} - \frac{1}{7^{2k+1}} + \cdots + \frac{(-1)^{n+1}}{(2n-1)^{2k+1}} + \cdots = \frac{\pi^{2k+1}}{2^{2k+2}(2k)!}E_k\,.^{\dagger} \tag{7.60}$$

7.2.4.2 Bernoulli and Euler Numbers

1. First Definition of the Bernoulli Numbers The Bernoulli numbers B_k occur in the power series expansion of some special functions, e.g., in the trigonometric functions $\tan x$, $\cot x$ and $\csc x$, also in the hyperbolic functions $\tanh x$, $\coth x$, and $\operatorname{cosech} x$. The Bernoulli numbers B_k can be defined as follows

$$\frac{x}{e^x-1} = 1 - \frac{x}{2} + B_1\frac{x^2}{2!} - B_2\frac{x^4}{4!} \pm \cdots + (-1)^{n+1}B_n\frac{x^{2n}}{(2n)!} \pm \cdots \qquad (|x| < 2\pi) \tag{7.61}$$

and they can be calculated by the coefficient comparison method with respect to the powers of x. The first in this way calculated values are given in **Table 7.1**.

Table 7.1 The first Bernoulli numbers

k	B_k	k	B_k	k	B_k	k	B_k
1	$\frac{1}{6}$	4	$\frac{1}{30}$	7	$\frac{7}{6}$	10	$\frac{174\,611}{330}$
2	$\frac{1}{30}$	5	$\frac{5}{66}$	8	$\frac{3\,617}{510}$	11	$\frac{854\,513}{138}$
3	$\frac{1}{42}$	6	$\frac{691}{2\,730}$	9	$\frac{43\,867}{798}$		

2. Second Definition of Bernoulli Numbers Some authors define the Bernoulli numbers in the following way:

$$\frac{x}{e^x-1} = 1 + \overline{B_1}\frac{x}{1!} + \overline{B_2}\frac{x^2}{2!} + \cdots + \overline{B_{2n}}\frac{x^{2n}}{(2n)!} + \cdots \qquad (|x| < 2\pi). \tag{7.62}$$

So one gets the recursive formula

$$\overline{B_{k+1}} = (\overline{B} + 1)^{k+1} \qquad (k = 1, 2, 3, \ldots), \tag{7.63}$$

where after the application of the binomial theorem (see 1.1.6.4, **1.**, p. 12) it is to replace $\overline{B}^{\nu}$ by $\overline{B}_{\nu}$, i.e., the exponent becomes the index. The first few numbers are:

$$\overline{B_1} = -\frac{1}{2},\quad \overline{B_2} = \frac{1}{6},\quad \overline{B_4} = -\frac{1}{30},\quad \overline{B_6} = \frac{1}{42},$$

$$\overline{B_8} = -\frac{1}{30},\quad \overline{B_{10}} = \frac{5}{66},\quad \overline{B_{12}} = -\frac{691}{2730},\quad \overline{B_{14}} = \frac{7}{6}, \tag{7.64}$$

* B_k are the Bernoulli numbers
† E_k are the Euler numbers

$$\overline{B_{16}} = -\frac{3617}{510}, \ldots, \qquad \overline{B_3} = \overline{B_5} = \overline{B_7} = \cdots = 0.$$

The following relation is valid:

$$B_k = (-1)^{k+1}\overline{B_{2k}} \qquad (k = 1, 2, 3, \ldots). \tag{7.65}$$

3. First Definition of Euler Numbers The Euler numbers E_k occur in the power series expansion of some special functions, e.g., in the functions $\sec x$ and $\operatorname{sech} x$. The EULER numbers E_k can be defined as follows

$$\sec x = 1 + E_1\frac{x^2}{2!} + E_2\frac{x^4}{4!} + \cdots + E_n\frac{x^{2n}}{(2n)!} + \cdots \qquad (|x| < \frac{\pi}{2}) \tag{7.66}$$

and they can be calculated by coefficient comparison with respect to the powers of x. Their values are given in **Table 7.2**.

4. Second Definition of Euler Numbers Analogously to (7.63) the *Euler numbers* can be defined with the recursive formula

$$(\overline{E} + 1)^k + (\overline{E} - 1)^k = 0 \qquad (k = 1, 2, 3, \ldots), \tag{7.67}$$

where after the application of the binomial theorem it is to replace $\overline{E^\nu}$ by $\overline{E_\nu}$. For the first values one gets:

$$\begin{aligned} &\overline{E_2} = -1, \quad \overline{E_4} = 5, \quad \overline{E_6} = -61, \quad \overline{E_8} = 1\,385, \\ &\overline{E_{10}} = -50\,521, \quad \overline{E_{12}} = 2\,702\,765, \quad \overline{E_{14}} = -199\,360\,981, \\ &\overline{E_{16}} = 19\,391\,512\,145, \ldots, \qquad \overline{E_1} = \overline{E_3} = \overline{E_5} = \cdots = 0. \end{aligned} \tag{7.68}$$

The following relation is valid:

$$E_k = (-1)^k\overline{E_{2k}} \qquad (k = 1, 2, 3, \ldots). \tag{7.69}$$

Table 7.2 First Euler numbers

k	E_k	k	E_k
1	1	5	50 521
2	5	6	2 702 765
3	61	7	199 360 981
4	1 385		

5. Relation Between the Euler and Bernoulli Numbers The relation between the Euler and Bernoulli numbers is:

$$\overline{E_{2k}} = \frac{4^{2k+1}}{2k+1}\left(\overline{B_k} - \frac{1}{4}\right)^{2k+1} \qquad (k = 1, 2, \ldots)\,. \tag{7.70}$$

7.2.5 Estimation of the Remainder

7.2.5.1 Estimation with Majorant

In order to determine how well the n-th partial sum approximates the sum of a convergent series, the absolute value of the remainder

$$|S - S_n| = |R_n| = \left|\sum_{k=n+1}^{\infty} a_k\right| \le \sum_{k=n+1}^{\infty} |a_k| \tag{7.71}$$

of the series $\sum\limits_{k=1}^{\infty} a_k$ must be estimated. For this estimation one uses a *majorant* for $\sum\limits_{k=n+1}^{\infty} |a_k|$, usually a geometric series or another series which is easy to sum or to estimate.

■ Estimate the remainder of the series $e = \sum_{n=0}^{\infty} \frac{1}{n!}$. For the ratio $\frac{a_{m+1}}{a_m}$ of two subsequent terms of this series with $m \geq n+1$ holds: $\frac{a_{m+1}}{a_m} = \frac{m!}{(m+1)!} = \frac{1}{m+1} \leq \frac{1}{n+2} = q < 1$. So the remainder $R_n = \frac{1}{(n+1)!} + \frac{1}{(n+2)!} + \frac{1}{(n+3)!} + \cdots$ can be majorized by the geometric series (7.15) with the quotient $q = \frac{1}{n+2}$ and with the initial term $a = \frac{1}{(n+1)!}$, and it yields:

$$R_n < \frac{a}{1-q} = \frac{1}{(n+1)!}\frac{n+2}{n+1} < \frac{1}{n!}\frac{n+2}{n^2+2n} = \frac{1}{n \cdot n!}. \tag{7.72}$$

7.2.5.2 Alternating Convergent Series

For a convergent alternating series, whose terms tend to zero with monotone decreasing absolute values, the easy estimation for the remainder is (see 7.2.3.3, **1.**, p. 463):

$$|R_n| = |S - S_n| < |a_{n+1}|. \tag{7.73}$$

7.2.5.3 Special Series

For some special series, e.g., Taylor series, there are special formulas to estimate the remainder (see 7.3.3.3, p. 471).

7.3 Function Series

7.3.1 Definitions

1. Function Series is a series whose terms are functions of the same variable x:

$$f_1(x) + f_2(x) + \cdots + f_n(x) + \cdots = \sum_{n=1}^{\infty} f_n(x). \tag{7.74}$$

2. Partial Sum $S_n(x)$ is the sum of the first n terms of the series (7.74):

$$S_n(x) = f_1(x) + f_2(x) + \cdots + f_n(x) = \sum_{k=1}^{n} f_k(x). \tag{7.75}$$

3. Domain of Convergence of a function series (7.74) is the set of values of $x = a$ for which all the functions $f_n(x)$ are defined and the series of constant terms

$$f_1(a) + f_2(a) + \cdots + f_n(a) + \cdots = \sum_{n=1}^{\infty} f_n(a) \tag{7.76}$$

is convergent, i.e., for which the *limit of the partial sums* $S_n(a)$ exists:

$$\lim_{n\to\infty} S_n(a) = \lim_{n\to\infty} \sum_{k=1}^{n} f_k(a) = S(a). \tag{7.77}$$

4. The Sum of the Series (7.74) is the function $S(x)$, and one says that the series converges to the function $S(x)$. The values for $x = a$ are the points of convergence.

5. Remainder $R_n(x)$ is the difference between the sum $S(x)$ of a convergent function series and its partial sum $S_n(x)$:

$$R_n(x) = S(x) - S_n(x) = f_{n+1}(x) + f_{n+2}(x) + \cdots + f_{n+m}(x) + \cdots. \tag{7.78}$$

7.3.2 Uniform Convergence

7.3.2.1 Definition, Weierstrass Theorem

According to the definition of the limit of a sequence of numbers (see 7.1.2, p. 458 and 7.2.1.1, **2.**, p. 459) the series (7.74) converges for all values x of a domain to $S(x)$ if for an arbitrary $\varepsilon > 0$ there is an integer $N(x)$ such that $|S(x) - S_n(x)| < \varepsilon$ holds for every $n > N(x)$. For function series there are to be distinguished between two cases:

1. Uniformly Convergent Series

If there is a number N such that for every x in the domain of convergence of the series (7.74), $|S(x) - S_n(x)| < \varepsilon$ holds for every $n > N$, then the series is called *uniformly convergent* on this domain.

2. Non-Uniform Convergence of Series

If there is no such number N which holds for every value of x in the domain of convergence, i.e., there are such values of ε for which there is at least one x in the domain of convergence such that $|S(x) - S_n(x)| > \varepsilon$ holds for arbitrarily large values of n, then the series is *non-uniformly convergent*.

■ **A** : The series $1 + \frac{x}{1!} + \frac{x^2}{2!} + \cdots + \frac{x^n}{n!} + \cdots = e^x$ (7.79a)

(see Table 21.5, p. 1059) is convergent for every value of x. The convergence is uniform for every bounded interval of x, and for every $|x| < a$ using the remainder of the Maclaurin formula (see 7.3.3.3, **2.**, p. 472) the inequality

$$|S(x) - S_n(x)| = \left|\frac{x^{n+1}}{(n+1)!}e^{\Theta x}\right| < \frac{a^{n+1}}{(n+1)!}e^a \qquad (0 < \Theta < 1) \tag{7.79b}$$

is valid. Since $(n+1)!$ increases faster than a^{n+1} in n, the expression on the right-hand side of the inequality, which is independent of x, will be less than ε for sufficiently large n. The series is not uniformly convergent on the whole numerical axis: For any large n there will be a value of x such that $\left|\frac{x^{n+1}}{(n+1)!}e^{\Theta x}\right|$ is greater than a previously given ε.

■ **B** : The series $x + x(1-x) + x(1-x)^2 + \cdots + x(1-x)^n + \cdots$, (7.80a)

is convergent for every x in $[0, 1]$, because corresponding to the d'Alembert ratio test (see 7.2.2.2, p. 461)

$$\rho = \lim_{n\to\infty}\left|\frac{a_{n+1}}{a_n}\right| = |1-x| < 1 \text{ is valid for } 0 < x \le 1 \text{ (for } x = 0\ S = 0 \text{ holds).} \tag{7.80b}$$

The convergence is non-uniform, because

$$S(x) - S_n(x) = x[(1-x)^{n+1} + (1-x)^{n+2} + \cdots] = (1-x)^{n+1} \tag{7.80c}$$

is valid and for every n there is an x such that $(1-x)^{n+1}$ is close enough to 1, i.e., it is not smaller than ε. In the interval $a \le x \le 1$ with $0 < a < 1$ the series is uniformly convergent.

3. Weierstrass Criterion for Uniform Convergence

The series (7.81a) is uniformly convergent in a given domain if there is a convergent series of constant positive terms (7.81b) such that for every x in this domain the inequality (7.81c) is valid.

$$f_1(x) + f_2(x) + \cdots + f_n(x) + \cdots \tag{7.81a}$$

$$c_1 + c_2 + \cdots + c_n + \cdots \tag{7.81b}$$

$$|f_n(x)| \le c_n \tag{7.81c}$$

(7.81b) is called a majorant of the series (7.81a).

7.3.2.2 Properties of Uniformly Convergent Series

1. Continuity

If the functions $f_1(x), f_2(x), \cdots, f_n(x), \cdots$ are continuous in a domain and the series $f_1(x)+f_2(x)+\cdots+f_n(x)+\cdots$ is uniformly convergent in this domain, then the sum $S(x)$ is continuous in the same domain.

If the series is not uniformly convergent in a domain, then the sum $S(x)$ may have discontinuities in this domain.

■ **A:** The sum of the series (7.80a) is discontinuous: $S(x) = 0$ for $x = 0$ and $S(x) = 1$ for $x > 0$.

■ **B:** The sum of the series (7.79a) is a continuous function: The series is non-uniformly convergent on the whole numerical axis. But it is uniformly convergent in every finite interval.

2. Integration and Differentiation of Uniformly Convergent Series

In the domain $[a, b]$ of uniform convergence it is allowed to integrate the series term-by-term. It is also allowed to differentiate term-by-term if the result is a uniformly convergent series. That is:

$$\int_{x_0}^{x} \sum_{n=1}^{\infty} f_n(t)\,dt = \sum_{n=1}^{\infty} \int_{x_0}^{x} f_n(t)\,dt \quad \text{for } x_0, x \in [a, b], \tag{7.82a}$$

$$\left(\sum_{n=1}^{\infty} f_n(x)\right)' = \sum_{n=1}^{\infty} f_n'(x) \qquad \text{for } x \in [a, b]. \tag{7.82b}$$

7.3.3 Power series

7.3.3.1 Definition, Convergence

1. Definition

The most important function series are the *power series* of the form

$$a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n + \cdots = \sum_{n=0}^{\infty} a_n x^n \quad \text{or} \tag{7.83a}$$

$$a_0 + a_1(x - x_0) + a_2(x - x_0)^2 + \cdots + a_n(x - x_0)^n + \cdots = \sum_{n=0}^{\infty} a_n (x - x_0)^n, \tag{7.83b}$$

where the coefficients a_i and the center of expansion x_0 are constant numbers.

2. Absolute Convergence and Radius of Convergence

A power series is convergent either only for $x = x_0$ or for all values of x or there is a number $r > 0$, the radius of convergence, such that the series is absolutely convergent for $|x - x_0| < r$ and divergent for $|x - x_0| > r$ **(Fig. 7.1)**. The *radius of convergence* can be calculated by the formulas

$$r = \lim_{n\to\infty} \left| \frac{a_n}{a_{n+1}} \right| \quad \text{or} \quad r = \frac{1}{\lim\limits_{n\to\infty} \sqrt[n]{|a_n|}} \tag{7.84}$$

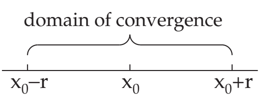


Figure 7.1

if the limits exist. If these limits do not exist, then one has to take the *limit superior* ($\overline{\lim}$) instead of the usual limit (see [7.6], Vol. I). At the endpoints $x = +r$ and $x = -r$ for the series (7.83a) and $x = x_0 + r$ and $x = x_0 - r$ for the series (7.83b) the series can be either convergent or divergent.

3. Uniform Convergence

A power series is uniformly convergent on every subinterval $|x - x_0| \le r_0 < r$ of the domain of convergence (*theorem of Abel*).

■ For the series $1 + \frac{x}{1} + \frac{x^2}{2} + \cdots + \frac{x^n}{n} + \cdots$ holds $\frac{1}{r} = \lim\limits_{n\to\infty} \frac{n+1}{n} = 1$, i.e., $r = 1$. (7.85)

Consequently the series is absolutely convergent in $-1 < x < +1$, for $x = -1$ it is conditionally convergent (see series (7.34), p. 462) and for $x = 1$ it is divergent (see the harmonic series (7.16), p. 459).

According to the theorem of Abel the series is uniformly convergent in every interval $[-r_1, +r_1]$, where r_1 is an arbitrary number between 0 and 1.

7.3.3.2 Calculations with Power Series

1. Sum and Product

Convergent power series can be added, multiplied, and multiplied by a constant factor term-by-term inside of their common domain of convergence. The product of two power series is

$$\left(\sum_{n=0}^{\infty} a_n x^n\right) \cdot \left(\sum_{n=0}^{\infty} b_n x^n\right) = a_0 b_0 + (a_0 b_1 + a_1 b_0)x + (a_0 b_2 + a_1 b_1 + a_2 b_0)x^2 + (a_0 b_3 + a_1 b_2 + a_2 b_1 + a_3 b_0)x^3 + \cdots . \tag{7.86}$$

2. First Terms of Some Powers of Power Series:

$$S = a + bx + cx^2 + dx^3 + ex^4 + fx^5 + \cdots , \tag{7.87}$$

$$S^2 = a^2 + 2abx + (b^2 + 2ac)x^2 + 2(ad + bc)x^3 + (c^2 + 2ae + 2bd)x^4 + 2(af + be + cd)x^5 + \cdots , \tag{7.88}$$

$$\sqrt{S} = S^{\frac{1}{2}} = a^{\frac{1}{2}} \left[1 + \frac{1}{2}\frac{b}{a}x + \left(\frac{1}{2}\frac{c}{a} - \frac{1}{8}\frac{b^2}{a^2}\right)x^2 + \left(\frac{1}{2}\frac{d}{a} - \frac{1}{4}\frac{bc}{a^2} + \frac{1}{16}\frac{b^3}{a^3}\right)x^3 + \left(\frac{1}{2}\frac{e}{a} - \frac{1}{4}\frac{bd}{a^2} - \frac{1}{8}\frac{c^2}{a^2} + \frac{3}{16}\frac{b^2c}{a^3} - \frac{5}{128}\frac{b^4}{a^4}\right)x^4 + \cdots\right] \quad (a > 0), \tag{7.89}$$

$$\frac{1}{\sqrt{S}} = S^{-\frac{1}{2}} = a^{-\frac{1}{2}} \left[1 - \frac{1}{2}\frac{b}{a}x + \left(\frac{3}{8}\frac{b^2}{a^2} - \frac{1}{2}\frac{c}{a}\right)x^2 + \left(\frac{3}{4}\frac{bc}{a^2} - \frac{1}{2}\frac{d}{a} - \frac{5}{16}\frac{b^3}{a^3}\right)x^3 + \left(\frac{3}{4}\frac{bd}{a^2} + \frac{3}{8}\frac{c^2}{a^2} - \frac{1}{2}\frac{e}{a} - \frac{15}{16}\frac{b^2c}{a^3} + \frac{35}{128}\frac{b^4}{a^4}\right)x^4 + \cdots\right] \quad (a > 0), \tag{7.90}$$

$$\frac{1}{S} = S^{-1} = a^{-1} \left[1 - \frac{b}{a}x + \left(\frac{b^2}{a^2} - \frac{c}{a}\right)x^2 + \left(\frac{2bc}{a^2} - \frac{d}{a} - \frac{b^3}{a^3}\right)x^3 + \left(\frac{2bd}{a^2} + \frac{c^2}{a^2} - \frac{e}{a} - 3\frac{b^2c}{a^3} + \frac{b^4}{a^4}\right)x^4 + \cdots\right] \quad (a \neq 0), \tag{7.91}$$

$$\frac{1}{S^2} = S^{-2} = a^{-2} \left[1 - 2\frac{b}{a}x + \left(3\frac{b^2}{a^2} - 2\frac{c}{a}\right)x^2 + \left(6\frac{bc}{a^2} - 2\frac{d}{a} - 4\frac{b^3}{a^3}\right)x^3 + \left(6\frac{bd}{a^2} + 3\frac{c^2}{a^2} - 2\frac{e}{a} - 12\frac{b^2c}{a^3} + 5\frac{b^4}{a^4}\right)x^4 + \cdots\right] \quad (a \neq 0). \tag{7.92}$$

3. Quotient of Two Power Series

$$\frac{\sum\limits_{n=0}^{\infty} a_n x^n}{\sum\limits_{n=0}^{\infty} b_n x^n} = \frac{a_0}{b_0}\frac{1 + \alpha_1 x + \alpha_2 x^2 + \cdots}{1 + \beta_1 x + \beta_2 x^2 + \cdots} = \frac{a_0}{b_0}[1 + (\alpha_1 - \beta_1)x + (\alpha_2 - \alpha_1\beta_1 + {\beta_1}^2 - \beta_2)x^2 + (\alpha_3 - \alpha_2\beta_1 - \alpha_1\beta_2 - \beta_3 - {\beta_1}^3 + \alpha_1{\beta_1}^2 + 2\beta_1\beta_2)x^3 + \cdots] \quad (b_0 \neq 0). \tag{7.93}$$

One gets this formula by considering the quotient (7.93) as a series with unknown coefficients, and after multiplying by the numerator the unknown coefficients follow by coefficient comparison.

4. Inverse of a Power Series

If the series

$$y = f(x) = ax + bx^2 + cx^3 + dx^4 + ex^5 + fx^6 + \cdots \quad (a \neq 0) \tag{7.94a}$$

is given, then its inverse is the series

$$x = \varphi(y) = Ay + By^2 + Cy^3 + Dy^4 + Ey^5 + Fy^6 + \cdots . \tag{7.94b}$$

Taking powers of y and comparing the coefficients yields

$$\begin{aligned} A &= \frac{1}{a}, \quad B = -\frac{b}{a^3}, \quad C = \frac{1}{a^5}(2b^2 - ac), \quad D = \frac{1}{a^7}(5abc - a^2d - 5b^3), \\ E &= \frac{1}{a^9}(6a^2bd + 3a^2c^2 + 14b^4 - a^3e - 21ab^2c), \\ F &= \frac{1}{a^{11}}(7a^3be + 7a^3cd + 84ab^3c - a^4f - 28a^2b^2d - 28a^2bc^2 - 42b^5). \end{aligned} \tag{7.94c}$$

The convergence of the inverse series must be checked in every case individually.

7.3.3.3 Taylor Series Expansion, Maclaurin Series

There is a collection of power series expansions of the most important elementary functions in **Table 21.5**, p. 1057. Usually, one gets them by Taylor expansion.

1. Taylor Series of Functions of One Variable

If a function $f(x)$ has all derivatives at $x = a$, then it can often be represented with the Taylor formula as a power series (see 6.1.4.5, p. 443).

a) First Form of the Representation:

$$f(x) = f(a) + \frac{x-a}{1!}f'(a) + \frac{(x-a)^2}{2!}f''(a) + \cdots + \frac{(x-a)^n}{n!}f^{(n)}(a) + \cdots . \tag{7.95a}$$

This representation (7.95a) is correct only for the x values for which the remainder $R_n = f(x) - S_n$ tends to zero if $n \to \infty$. This notion of the remainder is not identical to the notion of the remainder given in 7.3.1, p. 467 in general, only in the case if the expressions (7.95b) can be used.

There are the following formulas for the remainder:

$$R_n = \frac{(x-a)^{n+1}}{(n+1)!}f^{(n+1)}(\xi) \quad (a < \xi < x \text{ or } x < \xi < a) \quad \text{(Lagrange formula)}, \tag{7.95b}$$

$$R_n = \frac{1}{n!}\int_a^x (x-t)^n f^{(n+1)}(t)\,dt \qquad \text{(Integral formula)}. \tag{7.95c}$$

b) Second Form of the Representation:

$$f(a+h) = f(a) + \frac{h}{1!}f'(a) + \frac{h^2}{2!}f''(a) + \cdots + \frac{h^n}{n!}f^{(n)}(a) + \cdots . \tag{7.96a}$$

The expressions for the remainder are:

$$R_n = \frac{h^{n+1}}{(n+1)!}f^{(n+1)}(a + \Theta h) \qquad (0 < \Theta < 1), \tag{7.96b}$$

$$R_n = \frac{1}{n!}\int_0^h (h-t)^n f^{(n+1)}(a+t)\,dt. \tag{7.96c}$$

2. Maclaurin Series

The power series expansion of a function $f(x)$ is called a *Maclaurin series* if it is a special case of the Taylor series with $a = 0$. It has the form

$$f(x) = f(0) + \frac{x}{1!}f'(0) + \frac{x^2}{2!}f''(0) + \cdots + \frac{x^n}{n!}f^{(n)}(0) + \cdots \qquad (7.97a)$$

with the remainder

$$R_n = \frac{x^{n+1}}{(n+1)!}f^{(n+1)}(\Theta x) \quad (0 < \Theta < 1), \qquad (7.97b) \qquad R_n = \frac{1}{n!}\int_0^x (x-t)^n f^{(n+1)}(t)\,dt. \qquad (7.97c)$$

The convergence of the Taylor series and Maclaurin series can be proven either by examining the remainder R_n or determining the radius of convergence (see 7.3.3.1, p. 469). In the second case it can happen that although the series is convergent, the sum $S(x)$ is not equal to $f(x)$. This holds for instance for the function $f(x) = \exp(-\frac{1}{x^2})$ for $x \neq 0$, and $f(0) = 0$. The terms of its Maclaurin series are all identically zero, e.g., $S(x) = 0 \neq f(x)$.

7.3.4 Approximation Formulas

Considering only a neighborhood small enough of the center of the expansion, one can introduce rational approximation formulas for several functions with the help of the Taylor expansion. The first few terms of some functions are shown in **Table 7.3**. The data about accuracy are given by estimating the remainder. Further possibilities for approximate representation of functions, e.g., by interpolation and fitting polynomials or spline functions, can be found in 19.6, p. 982 and 19.7, p. 996. Important series expansions see **Table21.5**, p. 1057.

7.3.5 Asymptotic Power Series

Even divergent series can be useful for calculation of function values. Some asymptotic power series with respect to $\frac{1}{x}$ are considered next to calculate the values of functions for large values of $|x|$.

7.3.5.1 Asymptotic Behavior

Two functions $f(x)$ and $g(x)$, defined for $x_0 < x < \infty$, are called *asymptotically equal* for $x \to \infty$ if

$$\lim_{x\to\infty} \frac{f(x)}{g(x)} = 1 \qquad (7.98a) \qquad \text{or} \quad f(x) = g(x) + o(g(x)) \quad \text{for} \quad x \to \infty \qquad (7.98b)$$

hold. Here, $o(g(x))$ is the Landau symbol "little o" (see 2.1.4.9, p. 57). If (7.98a) or (7.98b) is fulfilled, then one writes also $f(x) \sim g(x)$.

■ **A:** $\sqrt{x^2+1} \sim x$. ■ **B:** $e^{\frac{1}{x}} \sim 1$. ■ **C:** $\frac{3x+2}{4x^3+x+2} \sim \frac{3}{4x^2}$.

7.3.5.2 Asymptotic Power Series

1. Notion of Asymptotic Series

A series $\sum_{\nu=0}^{\infty} \frac{a_\nu}{x^\nu}$ is called an *asymptotic power series* of the function $f(x)$ defined for $x > x_0$ if

$$f(x) = \sum_{\nu=0}^{n} \frac{a_\nu}{x^\nu} + O\left(\frac{1}{x^{n+1}}\right) \qquad (7.99)$$

holds for every $n = 0, 1, 2, \ldots$. Here, $O\left(\frac{1}{x^{n+1}}\right)$ is the Landau symbol "big O". For (7.99) one also writes $f(x) \approx \sum_{\nu=0}^{\infty} \frac{a_\nu}{x^\nu}$.

Table 7.3 Approximation formulas for some frequently used functions

Approximate formula	Next term	Tolerance interval for x with an error of					
		0.1%		1%		10%	
		from	to	from	to	from	to
$\sin x \approx x$	$-\frac{x^3}{6}$	-0.077 $-4.4°$	0.077 $4.4°$	-0.245 $-14.0°$	0.245 $14.0°$	-0.786 $-45.0°$	0.786 $45.0°$
$\sin x \approx x - \frac{x^3}{6}$	$+\frac{x^5}{120}$	-0.580 $-33.2°$	0.580 $33.2°$	-1.005 $-57.6°$	1.005 $57.6°$	-1.632 $-93.5°$	1.632 $93.5°$
$\cos x \approx 1$	$-\frac{x^2}{2}$	-0.045 $-2.6°$	0.045 $2.6°$	-0.141 $-8.1°$	0.141 $8.1°$	-0.415 $-25.8°$	0.415 $25.8°$
$\cos x \approx 1 - \frac{x^2}{2}$	$+\frac{x^4}{24}$	-0.386 $-22.1°$	0.386 $22.1°$	-0.662 $-37.9°$	0.662 $37.9°$	-1.036 $-59.3°$	1.036 $59.3°$
$\tan x \approx x$	$+\frac{x^3}{3}$	-0.054 $-3.1°$	0.054 $3.1°$	-0.172 $-9.8°$	0.172 $9.8°$	-0.517 $-29.6°$	0.517 $29.6°$
$\tan x \approx x + \frac{x^3}{3}$	$+\frac{2}{15}x^5$	-0.293 $-16.8°$	0.293 $16.8°$	-0.519 $-29.7°$	0.519 $29.7°$	-0.895 $-51.3°$	0.895 $51.3°$
$\sqrt{a^2+x} \approx a + \frac{x}{2a} = \frac{1}{2}\left(a + \frac{a^2+x}{a}\right)$	$-\frac{x^2}{8a^3}$	$-0.085a^2$	$0.093a^2$	$-0.247a^2$	$0.328a^2$	$-0.607a^2$	$1.545a^2$
$\frac{1}{\sqrt{a^2+x}} \approx \frac{1}{a} - \frac{x}{2a^3}$	$+\frac{3x^2}{8a^5}$	$-0.051a^2$	$0.052a^2$	$-0.157a^2$	$0.166a^2$	$-0.488a^2$	$0.530a^2$
$\frac{1}{a+x} \approx \frac{1}{a} - \frac{x}{a^2}$	$+\frac{x^2}{a^3}$	$-0.031a$	$0.031a$	$-0.099a$	$0.099a$	$-0.301a$	$0.301a$
$e^x \approx 1 + x$	$+\frac{x^2}{2}$	-0.045	0.045	-0.134	0.148	-0.375	0.502
$\ln(1+x) \approx x$	$-\frac{x^2}{2}$	-0.002	0.002	-0.020	0.020	-0.176	0.230

2. Properties of Asymptotic Power Series

a) Uniqueness: If for a function $f(x)$ the asymptotic power series exists, then it is unique, but a function is not uniquely determined by an asymptotic power series.

b) Convergence: Convergence is not required for an asymptotic power series.

■ **A:** $e^{\frac{1}{x}} \approx \sum_{\nu=0}^{\infty} \frac{1}{\nu! x^\nu}$ is an asymptotic series, which is convergent for every x with $|x| > x_0$ $(x_0 > 0)$.

■ **B:** The integral $f(x) = \int_0^\infty \frac{e^{-xt}}{1+t}\,dt$ $(x > 0)$ is convergent for $x > 0$ and repeated partial integration results in the representation $f(x) = \frac{1}{x} - \frac{1!}{x^2} + \frac{2!}{x^3} - \frac{3!}{x^4} \pm \cdots + (-1)^{n-1}\frac{(n-1)!}{x^n} + R_n(x)$ with $R_n(x) = (-1)^n \frac{n!}{x^n} \int_0^\infty \frac{e^{-xt}}{(1+t)^{n+1}}\,dt$. Because of $|R_n(x)| \le \frac{n!}{x^n}\int_0^\infty e^{-xt}\,dt = \frac{n!}{x^{n+1}}$ one gets

$R_n(x) = O\left(\frac{1}{x^{n+1}}\right)$, and with this estimation

$$\int_0^\infty \frac{e^{-xt}}{1+t}\,dt \approx \sum_{\nu=0}^{\infty}(-1)^\nu \frac{\nu!}{x^{\nu+1}}\,. \tag{7.100}$$

The asymptotic power series (7.100) is divergent for every x, because the absolute value of the quotient of the $(n+1)$-th and of the n-th terms has the value $\frac{n}{x}$. However, this divergent series can be used for a reasonably good approximation of $f(x)$. For instance, for $x = 10$ with the partial sums $S_4(10)$ and $S_5(10)$ the estimation $0.09152 < \int_0^\infty \frac{e^{-10t}}{1+t}\,dt < 0.09164$ holds.

7.4 Fourier Series

7.4.1 Trigonometric Sum and Fourier Series

7.4.1.1 Basic Notions

1. Fourier Representation of Periodic Functions

Often it is necessary or useful to represent a given periodic function $f(x)$ with period T exactly or approximatively by a sum of trigonometric functions of the form

$$\begin{aligned} s_n(x) = {} & \frac{a_0}{2} + a_1 \cos\omega x + a_2 \cos 2\omega x + \cdots + a_n \cos n\omega x \\ & + b_1 \sin\omega x + b_2 \sin 2\omega x + \cdots + b_n \sin n\omega x. \end{aligned} \tag{7.101}$$

This is called the *Fourier expansion.* Here the frequency is $\omega = \frac{2\pi}{T}$. In the case $T = 2\pi$ holds $\omega = 1$. One can get the best approximation of $f(x)$ in the sense given on 7.4.1.2, p. 475 by an approximation function $s_n(x)$, where the coefficients a_k and b_k $(k = 0, 1, 2, \ldots, n)$ are the Fourier coefficients of the given function. They are determined with the *Euler formulas*

$$a_k = \frac{2}{T}\int_0^T f(x)\cos k\omega x\,dx = \frac{2}{T}\int_{x_0}^{x_0+T} f(x)\cos k\omega x\,dx = \frac{2}{T}\int_0^{T/2} [f(x) + f(-x)]\cos k\omega x\,dx, \tag{7.102a}$$

and

$$b_k = \frac{2}{T}\int_0^T f(x)\sin k\omega x\,dx = \frac{2}{T}\int_{x_0}^{x_0+T} f(x)\sin k\omega x\,dx = \frac{2}{T}\int_0^{T/2} [f(x) - f(-x)]\sin k\omega x\,dx, \tag{7.102b}$$

or approximatively with the help of *methods of harmonic analysis* (see 19.6.4, p. 992).

2. Fourier Series

If there is a system of x values such that the sequence of functions $s_n(x)$ tends to a limit $s(x)$ for $n \to \infty$, then for the given function exists a *convergent Fourier series* for these x values. This can be written in the form

$$\begin{aligned} s(x) = {} & \frac{a_0}{2} + a_1 \cos\omega x + a_2 \cos 2\omega x + \cdots + a_n \cos n\omega x + \cdots \\ & + b_1 \sin\omega x + b_2 \sin 2\omega x + \cdots + b_n \sin n\omega x + \cdots \end{aligned} \tag{7.103a}$$

and also in the form

$$s(x) = \frac{a_0}{2} + A_1 \sin(\omega x + \varphi_1) + A_2 \sin(2\omega x + \varphi_2) + \cdots + A_n \sin(n\omega x + \varphi_n) + \cdots, \tag{7.103b}$$

where in the second case:

$$A_k = \sqrt{{a_k}^2 + {b_k}^2}, \quad \tan\varphi_k = \frac{a_k}{b_k}\,. \tag{7.103c}$$

3. Complex Representation of the Fourier Series

In many cases the complex form is very useful:

$$s(x) = \sum_{k=-\infty}^{+\infty} c_k e^{\mathrm{i}k\omega x}, \tag{7.104a}$$

$$c_k = \frac{1}{T}\int_0^T f(x)e^{-\mathrm{i}k\omega x}\,dx = \begin{cases} \frac{1}{2}a_0 & \text{for } k = 0\,, \\ \frac{1}{2}(a_k - \mathrm{i}b_k) & \text{for } k > 0\,, \\ \frac{1}{2}(a_{-k} + \mathrm{i}b_{-k}) & \text{for } k < 0\,. \end{cases} \tag{7.104b}$$

7.4.1.2 Most Important Properties of the Fourier Series

1. Least Mean Squares Error of a Function

If a function $f(x)$ over the interval $[0, T]$ $\left(T = \frac{2\pi}{\omega}\right)$ is approximated by a trigonometric sum

$$s_n(x) = \frac{a_0}{2} + \sum_{k=1}^{n} a_k \cos k\omega x + \sum_{k=1}^{n} b_k \sin k\omega x, \tag{7.105a}$$

also called the *Fourier sum*, then the mean square error (see 19.6.2.1, p. 984, and 19.6.4.1, **2.**, p. 993)

$$F = \frac{1}{T}\int_0^T [f(x) - s_n(x)]^2\,dx \tag{7.105b}$$

is smallest if one defines a_k and b_k as the Fourier coefficients (7.102a,b) of the given function.

2. Convergence of a Function in the Mean, Parseval Equation

The Fourier series converges over the interval $[0, T]$ $(T = \frac{2\pi}{\omega})$ in mean to the given function, i.e.,

$$\int_0^T [f(x) - s_n(x)]^2\,dx \to 0 \quad \text{for} \quad n \to \infty \tag{7.106a}$$

holds, if the function is bounded and in the interval $0 < x < T$ it is piecewise continuous. A consequence of the convergence in the mean is the *Parseval equation*:

$$\frac{2}{T}\int_0^T [f(x)]^2\,dx = \frac{a_0^2}{2} + \sum_{k=1}^{\infty}({a_k}^2 + {b_k}^2). \tag{7.106b}$$

3. Dirichlet Conditions

If the function $f(x)$ satisfies the *Dirichlet conditions*, i.e.,

a) the interval of definition can be decomposed into a finite number of intervals where the function $f(x)$ is continuous and monotone, and

b) at every point of discontinuity of $f(x)$ the values $f(x + 0)$ and $f(x - 0)$ are defined,

then the Fourier series of this function is convergent. At the points where $f(x)$ is continuous the sum is equal to $f(x)$, at the points of discontinuity the sum is equal to $\frac{f(x-0) + f(x+0)}{2}$.

4. Asymptotic Behavior of the Fourier Coefficients

If a periodic function $f(x)$ and its derivatives up to k-th order are continuous, then for $n \to \infty$ both expressions $a_n n^{k+1}$ and $b_n n^{k+1}$ tend to zero.

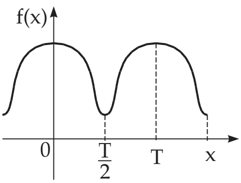


Figure 7.2

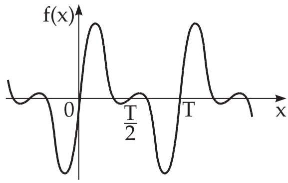


Figure 7.3

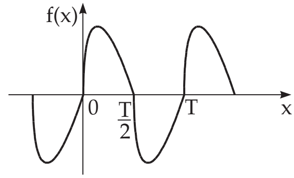


Figure 7.4

7.4.2 Determination of Coefficients for Symmetric Functions

7.4.2.1 Different Kinds of Symmetries

1. Symmetry of the First Kind

If the periodic function $f(x)$ with the period T is an even function, i.e., if $f(x) = f(-x)$ **(Fig. 7.2)**, then its Fourier coefficients are

$$a_k = \frac{4}{T}\int\limits_0^{T/2} f(x)\cos k\frac{2\pi x}{T}\,dx\,, \quad b_k = 0 \qquad (k = 0, 1, 2, \ldots)\,. \tag{7.107}$$

2. Symmetry of the Second Kind

If the periodic function $f(x)$ with the period T is an odd function, i.e., if $f(x) = -f(-x)$ **(Fig. 7.3)**, then its Fourier coefficients are

$$a_k = 0, \quad b_k = \frac{4}{T}\int\limits_0^{T/2} f(x)\sin k\frac{2\pi x}{T}\,dx \qquad (k = 0, 1, 2, \ldots). \tag{7.108}$$

3. Symmetry of the Third Kind

If for a periodic function with the period T holds $f(x + T/2) = -f(x)$ **(Fig. 7.4)**, then the Fourier coefficients are

$$a_{2k+1} = \frac{4}{T}\int\limits_0^{T/2} f(x)\cos(2k+1)\frac{2\pi x}{T}\,dx, \quad a_{2k} = 0, \tag{7.109a}$$

$$b_{2k+1} = \frac{4}{T}\int\limits_0^{T/2} f(x)\sin(2k+1)\frac{2\pi x}{T}\,dx, \quad b_{2k} = 0 \quad (k = 0, 1, 2, \ldots). \tag{7.109b}$$

4. Symmetry of the Fourth Kind

If the periodic function $f(x)$ with the period T is odd and also the symmetry of the third kind is satisfied **(Fig. 7.5a)**, then the Fourier coefficients are

$$a_k = b_{2k} = 0, \quad b_{2k+1} = \frac{8}{T}\int\limits_0^{T/4} f(x)\sin(2k+1)\frac{2\pi x}{T}\,dx \quad (k = 0, 1, 2, \ldots). \tag{7.110}$$

If the function $f(x)$ is even and also the symmetry of the third kind is satisfied **(Fig.7.5b)**, then the Fourier coefficients are

$$b_k = a_{2k} = 0, \quad a_{2k+1} = \frac{8}{T}\int\limits_0^{T/4} f(x)\cos(2k+1)\frac{2\pi x}{T}\,dx \quad (k = 0, 1, 2, \ldots). \tag{7.111}$$

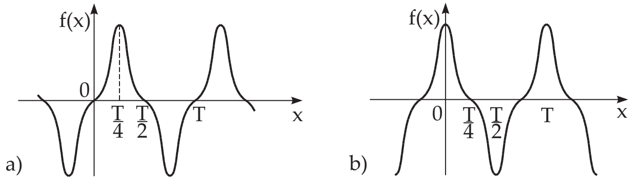


Figure 7.5

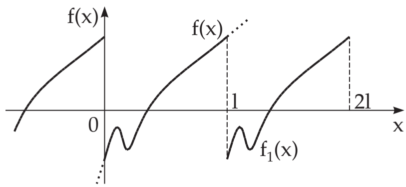


Figure 7.6

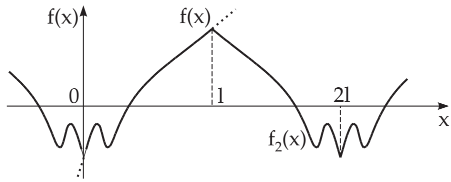


Figure 7.7

7.4.2.2 Forms of the Expansion into a Fourier Series

Every function $f(x)$, satisfying the Dirichlet conditions in an interval $0 \le x \le l$ (see 7.4.1.2, **3.**, p. 475), can be expanded in this interval into a convergent series of the following forms:

1. $$f_1(x) = \frac{a_0}{2} + a_1 \cos\frac{2\pi x}{l} + a_2 \cos 2\frac{2\pi x}{l} + \cdots + a_n \cos n\frac{2\pi x}{l} + \cdots$$
$$+\, b_1 \sin\frac{2\pi x}{l} + b_2 \sin 2\frac{2\pi x}{l} + \cdots + b_n \sin n\frac{2\pi x}{l} + \cdots. \tag{7.112a}$$

The period of the function $f_1(x)$ is $T = l$; in the interval $0 < x < l$ the function $f_1(x)$ coincides with the function $f(x)$ at the points of continuity **(Fig. 7.6)**. At the points of discontinuity $f_1(x) = \frac{1}{2}[f(x-0)+f(x+0)]$ holds. The coefficients of the expansion are determined with the Euler formulas (7.102a,b) for $\omega = \frac{2\pi}{l}$.

2. $$f_2(x) = \frac{a_0}{2} + a_1 \cos\frac{\pi x}{l} + a_2 \cos 2\frac{\pi x}{l} + \cdots + a_n \cos n\frac{\pi x}{l} + \cdots. \tag{7.112b}$$

The period of the function $f_2(x)$ is $T = 2l$; in the interval $0 \le x \le l$ the function $f_2(x)$ has a symmetry of the first kind and it coincides with the function $f(x)$ **(Fig. 7.7)**. The coefficients of the expansion of $f_2(x)$ are determined by the formulas for the case of symmetry of the first kind with $T = 2l$.

3. $$f_3(x) = b_1 \sin\frac{\pi x}{l} + b_2 \sin 2\frac{\pi x}{l} + \cdots + b_n \sin n\frac{\pi x}{l} + \cdots. \tag{7.112c}$$

The period of the function $f_3(x)$ is $T = 2l$; in the interval $0 < x < l$ the function $f_3(x)$ has a symmetry of the second kind and it coincides with the function $f(x)$ **(Fig. 7.8)**. The coefficients of the expansion are determined by the formulas for the case of symmetry of the second kind with $T = 2l$.

7.4.3 Determination of the Fourier Coefficients with Numerical Methods

If the periodic function $f(x)$ is a complicated one or in the interval $0 \le x < T$ its values are known only for a discrete system $x_k = \frac{kT}{N}$ with $k = 0, 1, 2, \ldots, N-1$, then the Fourier coefficients have to be

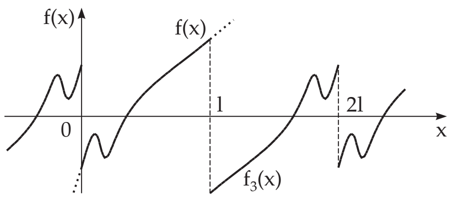


Figure 7.8

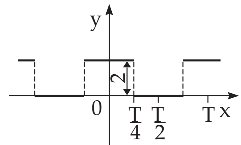


Figure 7.9

approximated. Furthermore, e.g., also the number of measurements N can be a very large number. In these cases one uses the methods of *numerical harmonic analysis* (see 19.6.4, p. 992).

7.4.4 Fourier Series and Fourier Integrals

1. Fourier integral

If the function $f(x)$ satisfies the Dirichlet conditions (see 7.4.1.2, **3.**, p. 475) in an arbitrarily finite interval and, moreover, the integral $\int\limits_{-\infty}^{+\infty} |f(x)|\,dx$ is convergent (see 8.2.3.2, **1.**, p. 507), then the following formula holds (FOURIER integral):

$$f(x) = \frac{1}{2\pi}\int\limits_{-\infty}^{+\infty} e^{\mathrm{i}\omega x}\,d\omega \int\limits_{-\infty}^{+\infty} f(t)e^{-\mathrm{i}\omega t}\,dt = \frac{1}{\pi}\int\limits_{0}^{\infty} d\omega \int\limits_{-\infty}^{+\infty} f(t)\cos\omega(t-x)\,dt. \tag{7.113a}$$

At the points of discontinuity can be used the substitution

$$f(x) = \frac{1}{2}[f(x-0) + f(x+0)]. \tag{7.113b}$$

2. Limiting Case of a Non-Periodic Function

The formula (7.113a) can be regarded as the expansion of a non-periodic function $f(x)$ into a trigonometric series in the interval $(-l, +l)$ for $l \to \infty$.

With Fourier series expansion a periodic function with period T is represented as the sum of harmonic vibrations with frequency $\omega_n = n\frac{2\pi}{T}$ with $n = 1, 2, \ldots$ and with amplitude A_n. This representation is based on a *discrete frequency spectrum*.

With the Fourier integral the non-periodic function $f(x)$ is represented as the sum of infinitely many harmonic vibrations with continuously varying frequency ω. The Fourier integral gives an expansion of the function $f(x)$ into a *continuous frequency spectrum*. Here the frequency ω corresponds to the density $g(\omega)$ of the spectrum:

$$g(\omega) = \frac{1}{2\pi}\int\limits_{-\infty}^{+\infty} f(t)e^{-\mathrm{i}\omega t}\,dt. \tag{7.113c}$$

The Fourier integral has a simpler form if $f(x)$ is either **a)** even or **b)** odd:

$$\textbf{a)}\quad f(x) = \frac{2}{\pi}\int\limits_{0}^{\infty} \cos\omega x\,d\omega \int\limits_{0}^{\infty} f(t)\cos\omega t\,dt, \tag{7.114a}$$

$$\textbf{b)}\quad f(x) = \frac{2}{\pi}\int\limits_{0}^{\infty} \sin\omega x\,d\omega \int\limits_{0}^{\infty} f(t)\sin\omega t\,dt. \tag{7.114b}$$

■ The density of the spectrum of the even function $f(x) = e^{-|x|}$ and the representation of this function are

$$g(\omega) = \frac{2}{\pi}\int_0^{\infty} e^{-t}\cos\omega t\,dt = \frac{2}{\pi}\frac{1}{\omega^2+1} \quad (7.115a) \qquad \text{and} \qquad e^{-|x|} = \frac{2}{\pi}\int_0^{\infty}\frac{\cos\omega x}{\omega^2+1}\,d\omega. \quad (7.115b)$$

7.4.5 Remarks on the Table of Some Fourier Expansions

In **Table 21.6**, p. 1062 there are given the Fourier expansions of some simple functions, which are defined in a certain interval and then they are periodically extended. The shapes of the curves of the expanded functions are graphically represented.

1. Application of Coordinate Transformations

Many of the simplest periodic functions can be reduced to a function represented in **Table 21.6** either by changing the scale (unit of measure) of the coordinate axis or by translation of the origin.

■ A function $f(x) = f(-x)$ defined by the relations

$$y = \begin{cases} 2 & \text{for } 0 < x < \dfrac{T}{4}, \\ 0 & \text{for } \dfrac{T}{4} < x < \dfrac{T}{2} \end{cases} \tag{7.116a}$$

(**Fig. 7.9**), can be transformed into the form 5 given in **Table 21.6**, by substituting $a = 1$ and introducing the new variables $Y = y - 1$ and $X = \dfrac{2\pi x}{T} + \dfrac{\pi}{2}$. By the substitution of the variables in series 5, because $\sin(2n+1)\left(\dfrac{2\pi x}{T} + \dfrac{\pi}{2}\right) = (-1)^n\cos(2n+1)\dfrac{2\pi x}{T}$ one gets for the function (7.116a) the expression

$$y = 1 + \frac{4}{\pi}\left(\cos\frac{2\pi x}{T} - \frac{1}{3}\cos 3\frac{2\pi x}{T} + \frac{1}{5}\cos 5\frac{2\pi x}{T} - \cdots\right). \tag{7.116b}$$

2. Using the Series Expansion of Complex Functions

Many of the formulas given in **Table 21.6** for the expansion of functions into trigonometric series can be derived from power series expansion of functions of a complex variable.

■ The expansion of the function

$$\frac{1}{1-z} = 1 + z + z^2 + \cdots \qquad (|z| < 1) \tag{7.117}$$

yields for

$$z = ae^{i\varphi} \tag{7.118}$$

after separating the real and imaginary parts

$$1 + a\cos\varphi + a^2\cos 2\varphi + \cdots + a^n\cos n\varphi + \cdots = \frac{1 - a\cos\varphi}{1 - 2a\cos\varphi + a^2},$$

$$a\sin\varphi + a^2\sin 2\varphi + \cdots + a^n\sin n\varphi + \cdots = \frac{a\sin\varphi}{1 - 2a\cos\varphi + a^2} \quad \text{for } |a| < 1. \tag{7.119}$$

8 Integral Calculus

1. Integral Calculus and Indefinite Integrals Integration represents the inverse operation of differentiation in the following sense: While differentiation calculates the derivative function $f'(x)$ of a given function $f(x)$, integration determines a function whose derivative $f'(x)$ is previously given. This process does not have a unique result, so it leads to the notion of an *indefinite integral*.

2. Definite Integral By starting with the graphical problem of the integral calculus, determining the area between the curve of $y = f(x)$ and the x-axis, and for this purpose approximating it with a set of thin rectangles **(Fig. 8.1)**, the notion of the *definite integral* can be introduced.

3. Connection Between Definite and Indefinite Integrals The relation between these two types of integral is the *fundamental theorem of calculus* (see 8.2.1.2, **1.**, p. 495).

8.1 Indefinite Integrals

8.1.1 Primitive Function or Antiderivative

1. Definition

Consider a function $y = f(x)$ given on an interval $[a, b]$. $F(x)$ is called a *primitive function* or *antiderivative* of $f(x)$ if $F(x)$ is differentiable everywhere on $[a, b]$ and its derivative is $f(x)$:

$$F'(x) = f(x). \tag{8.1}$$

Because under differentiation of $F(x) + C$ (C const) the additive constant disappears, a function has infinitely many primitive functions, if it has any. The difference of two primitive function is a constant. So, the graphs of all primitive functions $F_1(x), F_2(x), \ldots, F_n(x)$ can be got by parallel translation of a particular primitive function in the direction of the ordinate axis **(Fig. 8.2)**.

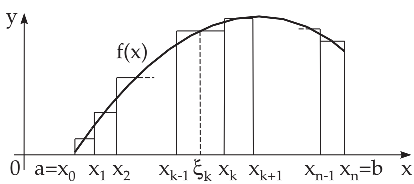


Figure 8.1

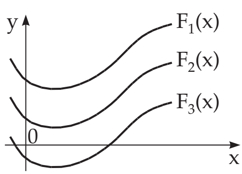


Figure 8.2

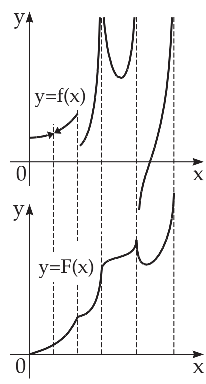


Figure 8.3

2. Existence

Every function continuous on an interval $[a, b]$ has a primitive function on this interval. If there are some discontinuities, then one decomposes the interval into subintervals in which the original function is continuous **(Fig. 8.3)**. The given function $y = f(x)$ is in the upper part of the figure; the function

I.N. Bronshtein et al., *Handbook of Mathematics*,
DOI 10.1007/978-3-662-46221-8_8

$y = F(x)$ in the lower part is a primitive function of it on the considered intervals.

8.1.1.1 Indefinite Integrals

The *indefinite integral* of a given function $f(x)$ is the set of primitive functions

$$F(x) + C = \int f(x)\,dx. \tag{8.2}$$

The function $f(x)$ under the integral sign $\int$ is called the *integrand*, x is the *integration variable*, and C is the *integration constant*. It is also a usual notation, especially in physics, to put the differential dx right after the integral sign and so before the function $f(x)$.

Table 8.1 Basic Integrals

Powers	**Exponential Functions**
$\int x^n\,dx = \frac{x^{n+1}}{n+1} \quad (n \neq -1)$ $\int \frac{dx}{x} = \ln\|x\|$	$\int e^x\,dx = e^x$ $\int a^x\,dx = \frac{a^x}{\ln a} \quad (a > 0, a \neq 1)$
Trigonometric Functions	**Hyperbolic Functions**
$\int \sin x\,dx = -\cos x$ $\int \cos x\,dx = \sin x$ $\int \tan x\,dx = -\ln\|\cos x\|$ $\int \cot x\,dx = \ln\|\sin x\|$ $\int \frac{dx}{\cos^2 x} = \tan x$ $\int \frac{dx}{\sin^2 x} = -\cot x$	$\int \sinh x\,dx = \cosh x$ $\int \cosh x\,dx = \sinh x$ $\int \tanh x\,dx = \ln\|\cosh x\|$ $\int \coth x\,dx = \ln\|\sinh x\|$ $\int \frac{dx}{\cosh^2 x} = \tanh x$ $\int \frac{dx}{\sinh^2 x} = -\coth x$
Fractional Rational Functions	**Irrational Functions**
$\int \frac{dx}{a^2+x^2} = \frac{1}{a}\arctan\frac{x}{a}$ $\int \frac{dx}{a^2-x^2} = \frac{1}{a}\operatorname{Artanh}\frac{x}{a} = \frac{1}{2a}\ln\left\|\frac{a+x}{a-x}\right\| \quad (\|x\| < a, a > 0)$ $\int \frac{dx}{x^2-a^2} = -\frac{1}{a}\operatorname{Arcoth}\frac{x}{a} = \frac{1}{2a}\ln\left\|\frac{x-a}{x+a}\right\| \quad (\|x\| > a, a > 0)$	$\int \frac{dx}{\sqrt{a^2-x^2}} = \arcsin\frac{x}{a} \quad (\|x\| < a, a > 0)$ $\int \frac{dx}{\sqrt{a^2+x^2}} = \operatorname{Arsinh}\frac{x}{a} = \ln\left\|x+\sqrt{x^2+a^2}\right\| \quad (a > 0)$ $\int \frac{dx}{\sqrt{x^2-a^2}} = \operatorname{Arcosh}\frac{x}{a} = \ln\left\|x+\sqrt{x^2-a^2}\right\| \quad (\|x\| > a, a > 0)$

8.1.1.2 Integrals of Elementary Functions

1. Basic Integrals

The integration of elementary functions in analytic form is reduced to a sequence of basic integrals. These basic integrals can be got from the derivatives of well-known elementary functions, since indefinite integration means the determination of a primitive function $F(x)$ of the function $f(x)$.
The collection of integrals given in **Table 8.1** comes from reversing the differentiation formulas in **Table 6.1**, p. 434 (Derivatives of elementary functions). The integration constant C is omitted.

2. General Case

For the solution of integration problems, one tries to reduce the given integral by algebraic and trigonometric transformations, or by using the integration rules to basic integrals. The integration methods given in section 8.1.2 make it possible in many cases to integrate those functions which have an elementary primitive function. The results of some integrations are collected in **Table 21.7**, p. 1065 (Indefinite integrals). The following remarks are very useful in integration:

a) The integration constant is mostly omitted. Exceptions are some integrals, which in different forms can be represented with different arbitrary constants.

b) If in the primitive function there is an expression containing $\ln f(x)$, then one has to consider always $\ln|f(x)|$ instead of it.

c) If the primitive function is given by a power series, then the function cannot be integrated in an elementary fashion.

A wide collection of integrals and their solutions are given in [8.1] and [8.2].

8.1.2 Rules of Integration

The integral of an integrand of arbitrary elementary functions is not usually an elementary function. In some special cases it is possible to use some tricks, and by practice one can gain some knowledge of how to integrate. Today the calculation of integrals is mostly left to computers.

The most important rules of integration, which are finally discussed here, are collected in **Table 8.2**.

1. Integrand with a Constant Factor

A constant factor α in the integrand can be factored out in front of the integral sign (*constant multiple rule*):

$$\int \alpha f(x)\,dx = \alpha \int f(x)\,dx. \tag{8.3}$$

2. Integration of a Sum or Difference

The integral of a sum or difference can be reduced to the integration of the separate terms if one can tell their integrals separately (*sum rule*):

$$\int (u+v-w)\,dx = \int u\,dx + \int v\,dx - \int w\,dx. \tag{8.4}$$

The variables u, v, w are functions of x.

■ $\int (x+3)^2(x^2+1)\,dx = \int (x^4+6x^3+10x^2+6x+9)\,dx = \frac{x^5}{5} + \frac{3}{2}x^4 + \frac{10}{3}x^3 + 3x^2 + 9x + C.$

3. Transformation of the Integrand

The integration of a complicated integrand can sometimes be reduced to a simpler integral by algebraic or trigonometric transformations.

■ $\int \sin 2x \cos x\,dx = \int \frac{1}{2}(\sin 3x + \sin x)\,dx.$

4. Linear Transformation in the Argument

If $\int f(x)\,dx = F(x)$ is known, e.g., from an integral table, then one gets:

$$\int f(ax)\,dx = \frac{1}{a}F(ax) + C, \tag{8.5a}$$

$$\int f(x+b)\,dx = F(x+b) + C, \tag{8.5b}$$

$$\int f(ax+b)\,dx = \frac{1}{a}F(ax+b) + C. \tag{8.5c}$$

■ **A:** $\int \sin ax\,dx = -\frac{1}{a}\cos ax + C.$ ■ **B:** $\int e^{ax+b}\,dx = \frac{1}{a}e^{ax+b} + C.$

■ **C:** $\displaystyle\int \frac{dx}{1+(x+a)^2} = \arctan(x+a) + C.$

Table 8.2 Important Rules of Calculation of Indefinite Integrals

Rule	**Formula for Integration**
Integration constant	$\int f(x)\,dx = F(x) + C \qquad (C \text{ const})$
Integration and differentiation	$F'(x) = \frac{dF}{dx} = f(x)$
Constant multiple rule	$\int \alpha f(x)\,dx = \alpha \int f(x)\,dx \qquad (\alpha \text{ const})$
Sum rule	$\int [u(x) \pm v(x)]\,dx = \int u(x)\,dx \pm \int v(x)\,dx$
Partial integration	$\int u(x)v'(x)\,dx = u(x)v(x) - \int u'(x)v(x)\,dx$
Substitution rule	$x = u(t)$ or $t = v(x)$; u und v are inverse functions of each other : $\int f(x)\,dx = \int f(u(t))u'(t)\,dt$ or $\int f(x)\,dx = \int \frac{f(u(t))}{v'(u(t))}\,dt$
Special form of the integrand	1. $\int \frac{f'(x)}{f(x)}\,dx = \ln\lvert f(x)\rvert + C$ (logarithmic integration) 2. $\int f'(x)f(x)\,dx = \frac{1}{2}f^2(x) + C$
Integration of the inverse function	u und v are inverse functions of each other : $\int u(x)\,dx = xu(x) - F(u(x)) + C_1$ with $F(x) = \int v(x)\,dx + C_2 \quad (C_1,\ C_2 \text{ const})$

5. Power and Logarithmic Integration

a) If the integrand has the form of a fraction such that the numerator is the derivative of the denominator, then the integral is the logarithm of the absolute value of the denominator:

$$\int \frac{f'(x)}{f(x)}\,dx = \int \frac{d\,f(x)}{f(x)} = \ln|f(x)| + C. \tag{8.6}$$

■ **A:** $\displaystyle\int \frac{2x+3}{x^2+3x-5}\,dx = \ln|x^2+3x-5| + C.$

b) If the integrand is a product of a power of a function multiplied by the derivative of the function, and the power is not equal to -1, then holds

$$\int f'(x)f^{\alpha}(x)\,dx = \int f^{\alpha}(x)d\,f(x) = \frac{f^{\alpha+1}(x)}{\alpha+1} + C \quad (\alpha \text{ const},\ \alpha \neq -1). \tag{8.7}$$

■ **B:** $\displaystyle\int \frac{2x+3}{(x^2+3x-5)^3}\,dx = \frac{1}{(-2)(x^2+3x-5)^2} + C.$

6. Substitution Method

If $x = u(t)$ where $t = v(x)$ is the inverse function of $x = u(t)$, then according to the chain rule of differentiation follows

$$\int f(x)\,dx = \int f[u(t)]u'(t)\,dt \quad \text{or} \quad \int f(x)\,dx = \int \frac{f(u(t))}{v'(u(t))}\,dt. \tag{8.8}$$

■ **A:** $\int \frac{e^x - 1}{e^x + 1}\,dx$. Substituting $x = \ln t$, $(t > 0)$, $\frac{dx}{dt} = \frac{1}{t}$, then taking the decomposition into partial fractions gives:
$\int \frac{e^x - 1}{e^x + 1}\,dx = \int \frac{t-1}{t+1}\frac{dt}{t} = \int \left(\frac{2}{t+1} - \frac{1}{t}\right)\,dt = 2\ln(e^x + 1) - x + C.$

■ **B:** $\int \frac{x\,dx}{1+x^2}$. Substituting $1 + x^2 = t$, $\frac{dt}{dx} = 2x$, then one gets $\int \frac{x\,dx}{1+x^2} = \int \frac{dt}{2t} = \frac{1}{2}\ln(1+x^2) + C$.

7. Partial Integration

Reversing the rule for the differentiation of a product gives

$$\int u(x)v^{'}(x)\,dx = u(x)\,v(x) - \int u^{'}(x)\,v(x)\,dx, \tag{8.9}$$

where $u(x)$ and $v(x)$ have continuous derivatives.

■ The integral $\int x\,e^x\,dx$ can be calculated by partial integration choosing $u = x$ and $v' = e^x$. Then $u^{'} = 1$ and $v = e^x$, therefore $\int xe^x\,dx = x\,e^x - \int e^x\,dx = (x-1)e^x + C$.

8. Non-Elementary Integrals

Integrals of elementary functions are not always elementary functions. These integrals are calculated mostly in the following three ways, where the primitive function will be approximated by a given accuracy:

1. Table of Values The integrals which have a particular theoretical or practical importance but cannot be expressed by elementary functions can be given by a table of values. (Of course, the table lists values of one particular primitive function.) Such special functions usually have special names. Examples are:

■ **A: Integral Logarithm** (see 8.2.5, **3.**, p. 513):

$$\int_0^x \frac{dt}{\ln t} = \mathrm{Li}\,(x). \tag{8.10}$$

■ **B: Elliptic integral of the first kind** (see 8.1.4.3, p. 490):

$$\int_0^{\sin\varphi} \frac{dx}{\sqrt{(1-x^2)(1-k^2x^2)}} = F(k,\varphi). \tag{8.11}$$

■ **C: Error function** (see 8.2.5, **5.**, p. 514):

$$\frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}\,dt = \mathrm{erf}(x). \tag{8.12}$$

2. Integration by Series Expansion Using the series expansion of the integrand, and if it is uniformly convergent, then it can be integrated term-by-term.

■ **A:** $\int \frac{\sin x}{x}\,dx$, (see also Sine integral p. 513).

■ **B:** $\int \frac{e^x}{x}\,dx$, (see also Exponential integral p. 514).

3. Graphical integration is the third approximation method, which is discussed in 8.2.1.4, **5.**, p. 499.

8.1.3 Integration of Rational Functions

Integrals of rational functions can always be expressed by elementary functions.

8.1.3.1 Integrals of Integer Rational Functions (Polynomials)

Integrals of integer rational functions are calculated directly by term-by-term integration:

$$\int (\, a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0)\, dx$$
$$= \frac{a_n}{n+1}x^{n+1} + \frac{a_{n-1}}{n}x^n + \cdots + \frac{a_1}{2}x^2 + a_0 x + C. \qquad (8.13)$$

8.1.3.2 Integrals of Fractional Rational Functions

The integrand of an integral of a fractional rational function $\int \frac{P(x)}{Q(x)}\,dx$, where $P(x)$ and $Q(x)$ are polynomials with degree m and n, respectively, can be transformed algebraically into a form which is easy to integrate. There are the following steps:

1. Simplifying the fraction by the greatest common divisor, so $P(x)$ and $Q(x)$ have no common factor.

2. Separating the integer and rational parts of the expression. If $m \geq n$, then $P(x)$ is divided by $Q(x)$. Then a polynomial and a proper fraction should be integrated.

3. Decomposing the denominator $Q(x)$ into linear and quadratic factors (see 1.6.3.2, p. 44):

$$Q(x) = a_n(x-\alpha)^k(x-\beta)^l \cdots (x^2+px+q)^r (x^2+p'x+q')^s \cdots \qquad (8.14a)$$

with $\frac{p^2}{4} - q < 0, \quad \frac{p'^2}{4} - q' < 0, \ldots .$ (8.14b)

4. Factoring out the constant coefficient a_n in front of the integral sign.

5. Decomposing the fraction into a sum of partial fractions: The proper fraction which is obtained after the division, cannot be further simplified and its denominator is decomposed into a product of irreducible factors. Then it can be decomposed into a sum of partial fractions (see 1.1.7.3, p. 15), each of them is easy to integrate.

8.1.3.3 Four Cases of Partial Fraction Decomposition

1. Case: All Roots of the Denominator are Real and Single

$$Q(x) = (x-\alpha)(x-\beta)\cdots(x-\lambda) \qquad (8.15a)$$

a) Forming the decomposition:

$$\frac{P(x)}{Q(x)} = \frac{A}{x-\alpha} + \frac{B}{x-\beta} + \cdots + \frac{L}{x-\lambda} \qquad (8.15b)$$

with $A = \frac{P(\alpha)}{Q'(\alpha)}, \quad B = \frac{P(\beta)}{Q'(\beta)}, \ldots, L = \frac{P(\lambda)}{Q'(\lambda)}\,.$ (8.15c)

b) The numbers $A, B, C, \ldots, L$ can also be calculated by the method of undetermined coefficients (see 1.1.7.3, **4.**, p. 17).

c) Integration by the formula

$$\int \frac{A\,dx}{x-\alpha} = A \ln|x-\alpha|. \qquad (8.15d)$$

■ $I = \int \frac{(2x+3)\,dx}{x^3+x^2-2x}$: $\frac{2x+3}{x(x-1)(x+2)} = \frac{A}{x} + \frac{B}{x-1} + \frac{C}{x+2}$, $A = \frac{P(0)}{Q'(0)} = \left(\frac{2x+3}{3x^2+2x-2}\right)_{x=0}$ $= -\frac{3}{2}$, $B = \left(\frac{2x+3}{3x^2+2x-2}\right)_{x=1} = \frac{5}{3}$, $C = \left(\frac{2x+3}{3x^2+2x-2}\right)_{x=-2} = -\frac{1}{6}$,

$I = \int \left(\frac{-\frac{3}{2}}{x} + \frac{\frac{5}{3}}{x-1} + \frac{-\frac{1}{6}}{x+2}\right) dx = -\frac{3}{2}\ln|x| + \frac{5}{3}\ln|x-1| - \frac{1}{6}\ln|x+2| + C_1 = \ln\left|\frac{C(x-1)^{5/3}}{x^{3/2}(x+2)^{1/6}}\right|.$

2. Case: All Roots of the Denominator are Real, Some of them with a Higher Multiplicity

$$Q(x) = (x-\alpha)^l (x-\beta)^m \cdots. \tag{8.16a}$$

a) Forming the decomposition:

$$\frac{P(x)}{Q(x)} = \frac{A_1}{(x-\alpha)} + \frac{A_2}{(x-\alpha)^2} + \cdots + \frac{A_l}{(x-\alpha)^l} + \frac{B_1}{(x-\beta)} + \frac{B_2}{(x-\beta)^2} + \cdots + \frac{B_m}{(x-\beta)^m} + \cdots. \tag{8.16b}$$

b) Calculation of the constants $A_1, A_2, \ldots, A_l, B_1, B_2, \ldots, B_m, \ldots$ by the method of undetermined coefficients (see 1.1.7.3, **4.**, p. 17).

c) Integration by the rule

$$\int \frac{A_1\,dx}{x-\alpha} = A_1 \ln|x-\alpha|, \quad \int \frac{A_k\,dx}{(x-\alpha)^k} = -\frac{A_k}{(k-1)(x-\alpha)^{k-1}} \quad (k>1). \tag{8.16c}$$

■ $I = \int \frac{x^3+1}{x(x-1)^3}\,dx$: $\frac{x^3+1}{x(x-1)^3} = \frac{A}{x} + \frac{B_1}{x-1} + \frac{B_2}{(x-1)^2} + \frac{B_3}{(x-1)^3}$. The method of undetermined coefficients yields $A + B_1 = 1$, $-3A - 2B_1 + B_2 = 0$, $3A + B_1 - B_2 + B_3 = 0$, $-A = 1$; $A = -1$, $B_1 = 2$, $B_2 = 1$, $B_3 = 2$. The result of the integration is

$I = \int \left[-\frac{1}{x} + \frac{2}{x-1} + \frac{1}{(x-1)^2} + \frac{2}{(x-1)^3}\right] dx = -\ln|x| + 2\ln|x-1| - \frac{1}{x-1} - \frac{1}{(x-1)^2} + C$
$= \ln\left|\frac{(x-1)^2}{x}\right| - \frac{x}{(x-1)^2} + C.$

3. Case: Some Roots of the Denominator are Single Complex

Suppose all coefficients of the denominator $Q(x)$ are real. Then, with a single complex root of $Q(x)$ its conjugate complex number is a root too and one can compose them into a quadratic polynomial.

$$Q(x) = (x-\alpha)^l (x-\beta)^m \ldots (x^2+px+q)(x^2+p'x+q')\ldots \tag{8.17a}$$

$$\text{with} \quad \frac{p^2}{4} < q, \quad \frac{p'^2}{4} < q', \ldots, \tag{8.17b}$$

because the quadratic polynomials have no real zeros.

a) Forming the decomposition:

$$\frac{P(x)}{Q(x)} = \frac{A_1}{x-\alpha} + \frac{A_2}{(x-\alpha)^2} + \cdots + \frac{A_l}{(x-\alpha)^l} + \frac{B_1}{x-\beta} + \frac{B_2}{(x-\beta)^2} + \cdots + \frac{B_m}{(x-\beta)^m} + \frac{Cx+D}{x^2+px+q} + \frac{Ex+F}{x^2+p'x+q'} + \cdots. \tag{8.17c}$$

b) Calculation of the constants by the method of undetermined coefficients (see 1.1.7.3, **4.**, p. 17).

c) Integration of the expression $\dfrac{Cx+D}{x^2+px+q}$ by the formula

$$\int \frac{(Cx+D)\,dx}{x^2+px+q} = \frac{C}{2}\ln|x^2+px+q| + \frac{D-Cp/2}{\sqrt{q-p^2/4}}\arctan\frac{x+p/2}{\sqrt{q-p^2/4}}. \tag{8.17d}$$

■ $I = \int \dfrac{4\,dx}{x^3+4x}$: $\quad \dfrac{4}{x^3+4x} = \dfrac{A}{x} + \dfrac{Cx+D}{x^2+4}$. The method of undetermined coefficients yields the equations $A+C=0,\ D=0,\ 4A=4,\ A=1,\ C=-1,\ D=0.$

$I = \int\left(\dfrac{1}{x} - \dfrac{x}{x^2+4}\right)dx = \ln|x| - \dfrac{1}{2}\ln(x^2+4) + \ln|C_1| = \ln\left|\dfrac{C_1 x}{\sqrt{x^2+4}}\right|$, where in this particular case the term arctan is missing.

4. Case: Some Roots of the Denominator are Complex with a Higher Multiplicity

$$Q(x) = (x-\alpha)^k(x-\beta)^l\ldots(x^2+px+q)^m(x^2+p'x+q')^n\ldots. \tag{8.18a}$$

a) Forming the decomposition:

$$\begin{aligned}\frac{P(x)}{Q(x)} &= \frac{A_1}{x-\alpha} + \frac{A_2}{(x-\alpha)^2} + \cdots + \frac{A_k}{(x-\alpha)^k} + \frac{B_1}{x-\beta} + \frac{B_2}{(x-\beta)^2} + \cdots + \frac{B_l}{(x-\beta)^l}\\ &\quad + \frac{C_1x+D_1}{x^2+px+q} + \frac{C_2x+D_2}{(x^2+px+q)^2} + \cdots + \frac{C_mx+D_m}{(x^2+px+q)^m}\\ &\quad + \frac{E_1x+F_1}{x^2+p'x+q'} + \frac{E_2x+F_2}{(x^2+p'x+q')^2} + \cdots + \frac{E_nx+F_n}{(x^2+p'x+q')^n}.\end{aligned} \tag{8.18b}$$

b) Calculation of the constants by the method of undetermined coefficients.

c) Integration of the expression $\dfrac{C_mx+D_m}{(x^2+px+q)^m}$ for $m>1$ in the following steps:

α) Transformation of the numerator into the form

$$C_mx+D_m = \frac{C_m}{2}(2x+p) + \left(D_m - \frac{C_mp}{2}\right). \tag{8.18c}$$

β) Decomposition of the integrand into the sum of two summands, where the first one can be integrated directly:

$$\int \frac{C_m}{2}\frac{(2x+p)\,dx}{(x^2+px+q)^m} = -\frac{C_m}{2(m-1)}\frac{1}{(x^2+px+q)^{m-1}}. \tag{8.18d}$$

γ) The second one will be integrated by the following recursion formula, not considering its coefficient:

$$\begin{aligned}\int \frac{dx}{(x^2+px+q)^m} &= \frac{x+p/2}{2(m-1)\,(q-p^2/4)\,(x^2+px+q)^{m-1}}\\ &\quad + \frac{2m-3}{2(m-1)\,(q-p^2/4)}\int\frac{dx}{(x^2+px+q)^{m-1}}.\end{aligned} \tag{8.18e}$$

■ $I = \int \dfrac{2x^2+2x+13}{(x-2)(x^2+1)^2}\,dx$: $\quad \dfrac{2x^2+2x+13}{(x-2)(x^2+1)^2} = \dfrac{A}{x-2} + \dfrac{C_1x+D_1}{x^2+1} + \dfrac{C_2x+D_2}{(x^2+1)^2}$.
The method of undetermined coefficients results in the following system of equations:
$A+C_1=0,\ -2C_1+D_1=0,\ 2A+C_1-2D_1+C_2=2,\ -2C_1+D_1-2C_2+D_2=2,\ A-2D_1-2D_2=13;$
the coefficients are $A=1,\ C_1=-1,\ D_1=-2,\ C_2=-3,\ D_2=-4,$

$I = \int\left(\dfrac{1}{x-2} - \dfrac{x+2}{x^2+1} - \dfrac{3x+4}{(x^2+1)^2}\right)dx.$

According to (8.18e) $\int \frac{dx}{(x^2+1)^2} = \frac{x}{2(x^2+1)} + \frac{1}{2}\int \frac{dx}{x^2+1} = \frac{x}{2(x^2+1)} + \frac{1}{2}\arctan x$ follows, and finally the result is $I = \frac{3-4x}{2(x^2+1)} + \frac{1}{2}\ln\frac{(x-2)^2}{x^2+1} - 4\arctan x + C$.

Table 8.3 Substitutions for Integration of Irrational Functions I

Integral *	**Substitution**
$\int R\left(x, \sqrt[n]{\frac{ax+b}{cx+e}}\right)\,dx$	$\sqrt[n]{\frac{ax+b}{cx+e}} = t$
$\int R\left(x, \sqrt[n]{\frac{ax+b}{cx+e}}, \sqrt[m]{\frac{ax+b}{cx+e}}\right)\,dx$	$\sqrt[r]{\frac{ax+b}{cx+e}} = t$ where r is the lowest common multiple of the numbers $m, n, \ldots$.
$\int R\left(x, \sqrt{ax^2+bx+c}\right)\,dx$:	One of the three *Euler substitutions*:
1. For $a > 0$ †	$\sqrt{ax^2+bx+c} = t - \sqrt{a}x$
2. For $c > 0$	$\sqrt{ax^2+bx+c} = xt + \sqrt{c}$
3. If the the polynomial ax^2+bx+c has different real roots: $ax^2+bx+c = a(x-\alpha)(x-\beta)$	$\sqrt{ax^2+bx+c} = t(x-\alpha)$

* The symbol R denotes a *rational* function of the expressions in parentheses. The numbers $n, m, \ldots$ are integers.
† If $a < 0$, and the polynomial ax^2+bx+c has complex roots, then the integrand is not defined for any value of x, since $\sqrt{ax^2+bx+c}$ is imaginary for every real value of x. In this case the integral is meaningless.

8.1.4 Integration of Irrational Functions

8.1.4.1 Substitution to Reduce to Integration of Rational Functions

Irrational functions cannot always be integrated in an elementary way. **Table 21.7**, p. 1065 contains a wide collection of integrals of irrational functions. In the simplest cases one can introduce substitutions, as in **Table 8.3**, such that the integral can be reduced to an integral of a rational function.

The integral $\int R\,(x, \sqrt{ax^2+bx+c})\,dx$ can be reduced to one of the following three forms

$$\int R\,(x, \sqrt{x^2+\alpha^2})\,dx, \qquad (8.19a)$$

$$\int R\,(x, \sqrt{x^2-\alpha^2})\,dx, \qquad (8.19b)$$

$$\int R\,(x, \sqrt{\alpha^2-x^2})\,dx, \qquad (8.19c)$$

because the quadratic polynomial ax^2+bx+c can always be written as the sum or as the difference of two complete squares. Then, one can use the substitutions given in **Table 8.4**.

■ **A:** $4x^2+16x+17 = 4\left(x^2+4x+4+\frac{1}{4}\right) = 4\left[(x+2)^2+\left(\frac{1}{2}\right)^2\right] = 4\left[x_1^2+\left(\frac{1}{2}\right)^2\right]$ with $x_1 = x+2$.

■ **B:** $x^2+3x+1=x^2+3x+\frac{9}{4}-\frac{5}{4}=\left(x+\frac{3}{2}\right)^2-\left(\frac{\sqrt{5}}{2}\right)^2=x_1^2-\left(\frac{\sqrt{5}}{2}\right)^2$ with $x_1=x+\frac{3}{2}$.

■ **C:** $-x^2+2x=1-x^2+2x-1=1^2-(x-1)^2=1^2-x_1^2$ with $x_1=x-1$.

Tabelle 8.4 Substitutions for Integration of Irrational Functions II

Integral	**Substitution**
$\int R\left(x,\sqrt{x^2+\alpha^2}\right)\,dx$	$x=\alpha\sinh t$ or $x=\alpha\tan t$
$\int R\left(x,\sqrt{x^2-\alpha^2}\right)\,dx$	$x=\alpha\cosh t$ or $x=\alpha\sec t$
$\int R\left(x,\sqrt{\alpha^2-x^2}\right)\,dx$	$x=\alpha\sin t$ or $x=\alpha\cos t$

8.1.4.2 Integration of Binomial Integrands

An expression of the form

$$x^m(a+bx^n)^p \tag{8.20}$$

is called a *binomial integrand*, where a and b are arbitrary real numbers, and m, n, p are arbitrary positive or negative rational numbers. The *theorem of Chebyshev* tells that the integral

$$\int x^m(a+bx^n)^p\,dx \tag{8.21}$$

can be expressed by elementary functions only in the following three cases:

Case 1: If p is an integer, then the expression $(a+bx^n)^p$ can be expanded by the binomial theorem, so the integrand after eliminating the parentheses will be a sum of terms in the form cx^k, which are easy to integrate.

Case 2: If $\frac{m+1}{n}$ is an integer, then the integral (8.21) can be reduced to the integral of a rational function by substituting $t=\sqrt[r]{a+bx^n}$, where r is the denominator of the fraction p.

Case 3: If $\frac{m+1}{n}+p$ is an integer, then the integral (8.21) can be reduced to the integral of a rational function by substituting $t=\sqrt[r]{\frac{a+bx^n}{x^n}}$, where r is the denominator of the fraction p.

■ **A:** $\int\frac{\sqrt[3]{1+\sqrt[4]{x}}}{\sqrt{x}}\,dx=\int x^{-1/2}\left(1+x^{1/4}\right)^{1/3}dx$; $m=-\frac{1}{2},\ n=\frac{1}{4},\ p=\frac{1}{3},\ \frac{m+1}{n}=2$, (Case 2):

Substitution $t=\sqrt[3]{1+\sqrt[4]{x}},\ x=(t^3-1)^4,\ dx=12t^2(t^3-1)^3dt;\ \int\frac{\sqrt[3]{1+\sqrt[4]{x}}}{\sqrt{x}}\,dx=12\int(t^6-t^3)\,dt$
$=\frac{3}{7}t^4(4t^3-7)+C.$

■ **B:** $\int\frac{x^3\,dx}{\sqrt[4]{1+x^3}}=\int x^3(1+x^3)^{-1/4}:\ m=3,\ n=3,\ p=-\frac{1}{4};\ \frac{m+1}{n}=\frac{4}{3},\ \frac{m+1}{n}+p=\frac{13}{12}.$

Because none of the three conditions is fulfilled, the integral is not an elementary function.

8.1.4.3 Elliptic Integrals

1. Indefinite Elliptic Integrals

Elliptic integrals are integrals of the form

$$\int R\,(x, \sqrt{ax^3+bx^2+cx+e})\,dx, \quad \int R\,(x, \sqrt{ax^4+bx^3+cx^2+ex+f})\,dx. \tag{8.22}$$

Usually they cannot be expressed by elementary functions; if it is still possible, the integral is called *pseudoelliptic*. The name of this type of integral originates from the fact that the first application of them was to calculate the perimeter of the ellipse (see 8.2.2.2, **2.**, p. 502). The inverses of elliptic integrals are the elliptic functions (see 14.6.1, p. 762). Integrals of the types (8.22), which are not integrable in elementary terms, can be reduced by a sequence of transformations into elementary functions and integrals of the following three types (see [21.1], [21.2], [21.7]):

$$\int \frac{dt}{\sqrt{(1-t^2)(1-k^2t^2)}} \quad (0<k<1), \quad (8.23a) \qquad \int \frac{(1-k^2t^2)\,dt}{\sqrt{(1-t^2)(1-k^2t^2)}} \quad (0<k<1), \quad (8.23b)$$

$$\int \frac{dt}{(1+nt^2)\sqrt{(1-t^2)(1-k^2t^2)}} \quad (0<k<1). \tag{8.23c}$$

Concerning the parameter n in (8.23c) one has to distinguish certain cases (see [14.1]).

By the substitution $t = \sin\varphi$ $\left(0<\varphi<\frac{\pi}{2}\right)$ the integrals (8.23a,b,c) can be transformed into the *Legendre form*:

Elliptic Integral of the First Kind: $$\int \frac{d\varphi}{\sqrt{1-k^2\sin^2\varphi}}. \tag{8.24a}$$

Elliptic Integral of the Second Kind: $$\int \sqrt{1-k^2\sin^2\varphi}\,d\varphi. \tag{8.24b}$$

Elliptic Integral of the Third Kind: $$\int \frac{d\varphi}{(1+n\sin^2\varphi)\sqrt{1-k^2\sin^2\varphi}}. \tag{8.24c}$$

2. Definite Elliptic Integrals

Definite integrals with zero as the lower bound corresponding to the indefinite elliptic integrals are denoted by

$$\int_0^{\varphi} \frac{d\psi}{\sqrt{1-k^2\sin^2\psi}} = F(k,\varphi), \quad (8.25a) \qquad \int_0^{\varphi} \sqrt{1-k^2\sin^2\psi}\,d\psi = E(k,\varphi), \quad (8.25b)$$

$$\int_0^{\varphi} \frac{d\psi}{(1+n\sin^2\psi)\sqrt{1-k^2\sin^2\psi}} = \Pi(n,k,\varphi) \quad \text{(for all three integrals } 0<k<1 \text{ holds).} \tag{8.25c}$$

These integrals are called *incomplete elliptic integrals* of the first, second, and third kind. For $\varphi = \frac{\pi}{2}$ the first two integrals are called *complete elliptic integrals*, and they are denoted by

$$K = F\left(k, \frac{\pi}{2}\right) = \int_0^{\frac{\pi}{2}} \frac{d\psi}{\sqrt{1-k^2\sin^2\psi}}, \tag{8.26a}$$

$$E = E\left(k, \frac{\pi}{2}\right) = \int_0^{\frac{\pi}{2}} \sqrt{1 - k^2 \sin^2 \psi}\, d\psi. \tag{8.26b}$$

Tables 21.9.1, 2, 3 contain the values for incomplete and complete elliptic integrals of the first and second kind F, E and also K and E. The series expansions of K and E see 8.2.5,**7.** p. 515.

■ The calculation of the perimeter of the ellipse leads to a complete elliptic integral of the second kind as a function of the numerical eccentricity e (see 8.2.2.2, **2.**, p. 502). For $a = 1.5,\ b = 1$ it follows that $e = 0.74$. Since $e = k = 0.74$ holds, one gets from **Table 21.9.3**, p. 1103: $\sin\alpha = 0.74$, i.e., $\alpha \approx 48°$ and $E(k, \frac{\pi}{2}) = E(0.74) = 1.3238$. It follows that $U = 4aE(0.74) \approx 4aE(\alpha = 48°) = 4\cdot 1.3238a \approx 7.94$. Numerical integration yields the better approximation 7.932 711 .

8.1.5 Integration of Trigonometric Functions

8.1.5.1 Substitution

With the *substitution*

$$t = \tan\frac{x}{2}, \quad \text{i.e.,} \quad dx = \frac{2\,dt}{1+t^2}, \quad \sin x = \frac{2t}{1+t^2}, \quad \cos x = \frac{1-t^2}{1+t^2}, \tag{8.27}$$

an integral of the form

$$\int R(\sin x, \cos x)\, dx \tag{8.28}$$

can be transformed into an integral of a rational function, where R denotes a rational function of its arguments.

■ $\displaystyle\int \frac{1+\sin x}{\sin x(1+\cos x)}\, dx = \int \frac{\left(1 + \frac{2t}{1+t^2}\right)\frac{2}{1+t^2}}{\frac{2t}{1+t^2}\left(1 + \frac{1-t^2}{1+t^2}\right)}\, dt = \frac{1}{2}\int\left(t + 2 + \frac{1}{t}\right) dt = \frac{t^2}{4} + t + \frac{1}{2}\ln|t| + C$

$\displaystyle = \frac{\tan^2\frac{x}{2}}{4} + \tan\frac{x}{2} + \frac{1}{2}\ln\left|\tan\frac{x}{2}\right| + C.$ In some special cases more simple substitutions can be applied.

If the integrand in (8.28) contains only odd powers of the functions $\sin x$ and $\cos x$, then by the substitution $t = \tan x$ a rational function can be obtained in a simpler way.

8.1.5.2 Simplified Methods

Case 1: $\displaystyle\int R(\sin x)\cos x\, dx.$ Substitution $t = \sin x, \quad \cos x\, dx = dt\,.$ (8.29)

Case 2: $\displaystyle\int R(\cos x)\sin x\, dx.$ Substitution $t = \cos x, \quad \sin x\, dx = -dt.$ (8.30)

Case 3: $\displaystyle\int \sin^n x\, dx:$ (8.31a)

a) $n = 2m+1$, odd:

$$\int \sin^n x\, dx = \int (1-\cos^2 x)^m \sin x\, dx = -\int (1-t^2)^m\, dt \qquad \text{with} \quad t = \cos x. \tag{8.31b}$$

b) $n = 2m$, even:

$$\int \sin^n x\, dx = \int \left[\frac{1}{2}(1-\cos 2x)\right]^m dx = \frac{1}{2^{m+1}}\int (1-\cos t)^m\, dt \qquad \text{with} \quad t = 2x. \tag{8.31c}$$

The power is halved in this way. After removing the parentheses in $(1 - \cos t)^m$ follows integration term-by-term.

Case 4: $\int \cos^n x\,dx.$ (8.32a)

a) $n = 2m + 1$, odd:

$$\int \cos^n x\,dx = \int (1 - \sin^2 x)^m \cos x\,dx = \int (1 - t^2)^m\,dt \qquad \text{with} \quad t = \sin x. \tag{8.32b}$$

b) $n = 2m$, even:

$$\int \cos^n x\,dx = \int \left[\frac{1}{2}(1 + \cos 2x)\right]^m dx = \frac{1}{2^{m+1}} \int (1 + \cos t)^m\,dt \qquad \text{with} \quad t = 2x. \tag{8.32c}$$

The power is halved in this way. After removing the parentheses follows integration term-by-term.

Case 5: $\int \sin^n x \cos^m x\,dx.$ (8.33a)

a) One of the numbers m or n is odd: Reducing it to the cases 1 or 2.

■ **A:** $\int \sin^2 x \cos^5 x\,dx = \int \sin^2 x\,(1 - \sin^2 x)^2 \cos x\,dx = \int t^2(1 - t^2)^2\,dt + C$ with $t = \sin x$.

■ **B:** $\int \frac{\sin x}{\sqrt{\cos x}} dx = -\int \frac{dt}{\sqrt{t}} + C$ with $t = \cos x$.

b) The numbers m and n are both even: Reducing it to the cases 3 or 4 by halving the powers using the trigonometric formulas

$$\sin x \cos x = \frac{\sin 2x}{2}, \qquad \sin^2 x = \frac{1 - \cos 2x}{2}, \qquad \cos^2 x = \frac{1 + \cos 2x}{2}. \tag{8.33b}$$

■ $\int \sin^2 x \cos^4 x\,dx = \int (\sin x \cos x)^2 \cos^2 x\,dx = \frac{1}{8} \int \sin^2 2x(1 + \cos 2x)\,dx = \frac{1}{8} \int \sin^2 2x \cos 2x\,dx + \frac{1}{16} \int (1 - \cos 4x)\,dx = \frac{1}{48} \sin^3 2x + \frac{1}{16} x - \frac{1}{64} \sin 4x + C.$

Case 6:

$$\int \tan^n x\,dx = \int \tan^{n-2} x(\sec^2 x - 1)\,dx = \int \tan^{n-2} x\,(\tan x)'\,dx - \int \tan^{n-2} x\,dx$$
$$= \frac{\tan^{n-1} x}{n - 1} - \int \tan^{n-2} x\,dx. \tag{8.34a}$$

By repeating this process one decreases the power and depending on whether n is even or odd one finally gets the integral

$$\int dx = x \qquad \text{or} \qquad \int \tan x\,dx = -\ln|\cos x| \tag{8.34b}$$

respectively.

Case 7: $\int \cot^n x\,dx.$ (8.35)

The solution is similar to case 6.

Remark: Table 21.7, p. 1065 contains several integrals with trigonometric functions.

8.1.6 Integration of Further Transcendental Functions

8.1.6.1 Integrals with Exponential Functions

Integrals with exponential functions can be reduced to integrals of rational functions if it is given in the form

$$\int R\,(e^{mx}, e^{nx}, \ldots, e^{px})\,dx, \tag{8.36a}$$

where $m, n, \ldots, p$ are rational numbers. Necessary are two substitutions to calculate the integral:
1. Substitution of $t = e^x$ results in an integral

$$\int \frac{1}{t} R(t^m, t^n, \ldots, t^p)\, dt. \tag{8.36b}$$

2. Substitution of $z = \sqrt[r]{t}$, where r is the lowest common multiple of the denominators of the fractions $m, n, \ldots, p$, results in an integral of a rational function.

8.1.6.2 Integrals with Hyperbolic Functions

Integrals with hyperbolic functions, i.e., containing the functions $\sinh x, \cosh x, \tanh x$ and $\coth x$ in the integrand, can be calculated as integrals with exponential functions, if the hyperbolic functions are replaced by the corresponding exponential functions. The most often occurring cases $\int \sinh^n x\, dx$, $\int \cosh^n x\, dx$, $\int \sinh^n x \cosh^m x\, dx$ can be integrated in a similar way to the trigonometric functions (see 8.1.5, . p. 491).

8.1.6.3 Application of Integration by Parts

If the integrand is a logarithm, inverse trigonometric function, inverse hyperbolic function or a product of x^m with $\ln x$, e^{ax}, $\sin ax$ or $\cos ax$ or their inverses, then the solution can be got by a single or repeated integration by parts.
In some cases the repeated partial integration results in an integral of the same type as the original integral. In this case one has to solve an algebraic equation with respect to this expression. One can calculate in this way, e.g., the integrals $\int e^{ax} \cos bx\, dx$, $\int e^{ax} \sin bx\, dx$, where one needs integration by parts twice. The same type of function should be chosen for the factor u in both steps, either the exponential or the trigonometric function.

One also uses integration by parts if there are integrals in the forms $\int P(x) e^{ax}\, dx$, $\int P(x) \sin bx\, dx$ and $\int P(x) \cos bx\, dx$, where $P(x)$ is a polynomial. (Choosing $u = P(x)$ the degree of the polynomial will be decreased at every step.)

8.1.6.4 Integrals of Transcendental Functions

The **Table 21.7**, p. 1065, contains many integrals of transcendental functions.

8.2 Definite Integrals

8.2.1 Basic Notions, Rules and Theorems

8.2.1.1 Definition and Existence of the Definite Integral

1. Definition of the Definite Integral

The definite integral of a bounded function $y = f(x)$ defined on a finite closed interval $[a, b]$ is a number, which is defined as a limit of a sum, where either $a < b$ can hold (case A) or $a > b$ can hold (case B).

In a generalization of the notion of the definite integral (see 8.2.3, p. 506) in the following sections also functions are considered, which are defined on an arbitrary connected domain of the real line, e.g., on an open or half-open interval, on a half-axis or on the whole numerical axis, or on a domain which is only piecewise connected, i.e., everywhere, except finitely many points. These types of integrals belong to *improper integrals* (see 8.2.3, **1.**, p. 506).

2. Definite Integral as the Limit of a Sum

The notion of the definite integral is obtained by the limit given in the following procedure (see **Fig. 8.1**, p. 480):

1. Step: The interval $[a, b]$ is decomposed into n *subintervals* by the choice of $n - 1$ arbitrary points $x_1, x_2, \ldots, x_{n-1}$ so that one of the following cases occurs:

$$a = x_0 < x_1 < x_2 < \cdots < x_i < \cdots < x_{n-1} < x_n = b \qquad \text{(case A)} \quad \text{or} \tag{8.37a}$$

$$a = x_0 > x_1 > x_2 > \cdots > x_i > \cdots > x_{n-1} > x_n = b \qquad \text{(case B)}. \tag{8.37b}$$

2. Step: A point ξ_i is chosen in the inside or on the boundary of each subinterval as in **Fig. 8.4**:

$$x_{i-1} \le \xi_i \le x_i \quad \text{(in case A)} \qquad \text{or} \qquad x_{i-1} \ge \xi_i \ge x_i \quad \text{(in case B)}. \tag{8.37c}$$

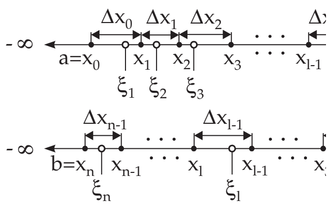


Figure 8.4

3. Step: The value $f(\xi_i)$ of the function $f(x)$ at the chosen point is multiplied by the corresponding difference $\Delta x_{i-1} = x_i - x_{i-1}$, i.e., by the length of the subinterval taken with a positive sign in case A and taken with negative sign in case B. This step is represented in **Fig. 8.1**, p. 480 for the case A.

4. Step: Then all the n products $f(\xi_i)\,\Delta x_{i-1}$ are added.

5. Step: The limit of the obtained *integral approximation sum* or *Riemann sum*

$$\sum_{i=1}^{n} f(\xi_i)\,\Delta x_{i-1} \tag{8.38}$$

is calculated if the length of each subinterval Δx_{i-1} tends to zero and consequently their number n tends to ∞. Based on this, one can also denote Δx_{i-1} as an infinitesimal quantity.

If this limit exists independently of the choice of the numbers x_i and ξ_i, then it is called the *definite Riemann integral* of the considered function on the given interval. One writes

$$\int_a^b f(x)\,dx = \lim_{\substack{\Delta x_{i-1} \to 0 \\ n \to \infty}} \sum_{i=1}^{n} f(\xi_i)\,\Delta x_{i-1}. \tag{8.39}$$

The endpoints of the interval are called *limits of integration* and the interval $[a, b]$ is the *integration interval*; a is the *lower limit*, b is the *upper limit of integration*; x is called the *integration variable* and $f(x)$ is called the *integrand*.

3. Existence of the Definite Integral

The definite integral of a continuous function on $[a, b]$ is always defined, i.e., the limit (8.39) always exists and is independent of the choice of the numbers x_i and ξ_i. Also for a bounded function having only a finite number of discontinuities on the interval $[a, b]$ the definite integral exists. The function whose definite integral exists on a given interval is called an *integrable function* on this interval.

8.2.1.2 Properties of Definite Integrals

The most important properties of definite integrals explained in the following section are presented in **Table 8.5**, p. 496.

1. Fundamental Theorem of Integral Calculus

If the integrand $f(x)$ is continuous on the interval $[a,b]$, and $F(x)$ is a primitive function, then

$$\int_a^b f(x)\,dx = \int_a^b F'(x)\,dx = F(x)|_a^b = F(b) - F(a) \tag{8.40}$$

holds, i.e., the calculation of a definite integral is reduced to the calculation of the corresponding indefinite integral, to the determination of the antiderivative:

$$F(x) = \int f(x)\,dx + C. \tag{8.41}$$

Remark: There are integrable functions which do not have any primitive function, but one will see that, if a function is continuous, it has a primitive function.

2. Geometric Interpretation and Rule of Signs

1. Area under a Curve Let $f(x) \geq 0$ for all x in $[a,b]$. Then the sum (8.38) can be considered as the total area of the rectangles **(Fig. 8.1)**, p. 480, which approximate the area under the curve $y = f(x)$. Therefore the limit of this sum and together with it the definite integral is equal to the area of the region A, which is bounded by the curve $y = f(x)$, the x-axis, and the parallel lines $x = a$ and $x = b$:

$$A = \int_a^b f(x)\,dx \quad (a < b \text{ and } f(x) \geq 0 \text{ for } a \leq x \leq b). \tag{8.42}$$

2. Sign Rule If a function $y = f(x)$ is piecewise positive or negative in the integration interval **(Fig. 8.5)**, then the integrals over the corresponding subintervals, that is, the area parts, have positive or negative values, so the integration over the total interval yields the sum of signed areas.
In **Fig. 8.5a–d** four cases are represented with the different possibilities of the sign of the area.

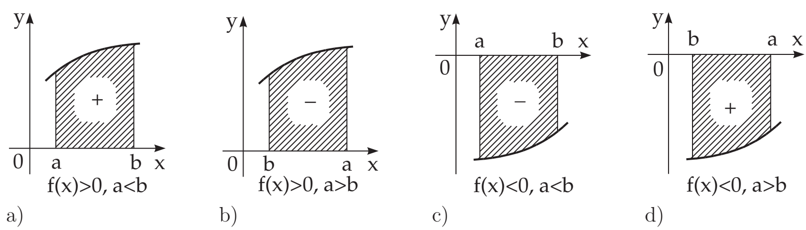


Figure 8.5

■ **A:** $\int_0^{\pi} \sin x\,dx$ (read: Integral from $x = 0$ to $x = \pi$) $= (-\cos x|_0^{\pi} = (-\cos\pi + \cos 0) = 2.$

■ **B:** $\int_0^{2\pi} \sin x\,dx$ (read: Integral from $x = 0$ to $x = 2\pi$) $= (-\cos|_0^{2\pi} = (-\cos 2\pi + \cos 0) = 0.$

3. Variable Upper Limit

1. Particular Integral If the upper limit is considered as variable (**Fig. 8.6**, region $ABCD$), then there is an area function in the form

$$S(x) = \int_a^x f(t)\,dt \quad (a < b \text{ and } f(x) \geq 0 \text{ for } x \geq a). \tag{8.43}$$

This integral is called a *particular integral*.

To avoid accidently confusing the variable upper limit x with the variable of the integrand, often the integration variable is denoted by t instead of x as in (8.43).

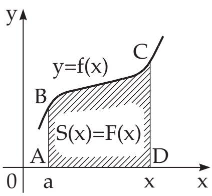


Figure 8.6

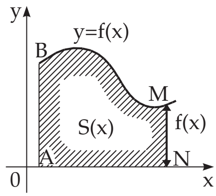


Figure 8.7

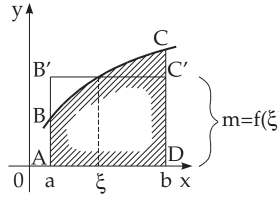


Figure 8.8

2. Differentiation of the Definite Integral with Respect to the Upper Limit A definite integral with a variable upper limit $\int_a^x f(t)\,dt$, if this integral exists, is a continuous function $F(x)$ of the upper limit. If $f(x)$ is continuous, then $F(x)$ is differentiable with respect to x, i.e., it is a primitive function of the integrand. So, if $f(x)$ is continuous on $[a, b]$, and $x \in (a, b)$ holds

$$F'(x) = f(x) \qquad \text{or} \qquad \frac{d}{dx}\int\limits_a^x f(t)\,dt = f(x). \tag{8.44}$$

The geometrical meaning of this theorem is that the derivative of the variable area $S(x)$ is equal to the length of the segment NM **(Fig. 8.7)**. Here, the area, just as the length of the segment, is considered according to the sign rule **(Fig. 8.5)**.

4. Decomposition of the Integration Interval

The interval of integration $[a, b]$ can be decomposed into subintervals. The value of the definite integral over the complete interval is

$$\int\limits_a^b f(x)\,dx = \int\limits_a^c f(x)\,dx + \int\limits_c^b f(x)\,dx. \tag{8.45}$$

This is called the *interval rule*. If the integrand has finitely many jumps, then the interval can be decomposed into subintervals such that on these subintervals the integrand will already be continuous. Then the integral can be calculated according to the above formula as the sum of the integrals on the subintervals.
At the endpoints of the subintervals the function must be defined by its corresponding left or right-sided limit, if it exists. If it does not, then the integral is an improper integral (see 8.2.3.3, **1.**, p. 509).

Remark: The formula above is valid also in the case if c is outside of the interval $[a, b]$ if one can suppose that the integrals on the right-hand side exist.

8.2.1.3 Further Theorems about the Limits of Integration

1. Independence of the Notation of the Integration Variable

The value of a definite integral is independent of the notation of the integration variable:

$$\int\limits_a^b f(x)\,dx = \int\limits_a^b f(u)\,du = \int\limits_a^b f(t)\,dt. \tag{8.46}$$

Table 8.5 Important Properties of Definite Integrals

Property	Formula
Fundamental theorem of the integral calculus ($f(x)$ is continuous)	$\int\limits_a^b f(x)\,dx = F(x)\Big\vert_a^b = F(b) - F(a)$ with $F(x) = \int f(x)\,dx + C$ or $F'(x) = f(x)$
Interchange rule	$\int\limits_a^b f(x)\,dx = -\int\limits_b^a f(x)\,dx$
Equal integration limits	$\int\limits_a^a f(x)\,dx = 0$
Interval rule	$\int\limits_a^b f(x)\,dx = \int\limits_a^c f(x)\,dx + \int\limits_c^b f(x)\,dx$
Independence of the notation of the integration variable	$\int\limits_a^b f(x)\,dx = \int\limits_a^b f(u)\,du = \int\limits_a^b f(t)\,dt$
Differentiation with respect to the upper limit	$\frac{d}{dx}\int\limits_a^x f(t)\,dt = f(x)$ with $f(x)$ continuous
Mean value theorem of the integral calculus	$\int\limits_a^b f(x)\,dx = (b-a)f(\xi) \quad (a < \xi < b)$

2. Equal Integration Limits

If the lower and upper limits are equal, then the value of the integral is equal to zero:

$$\int\limits_a^a f(x)\,dx = 0. \tag{8.47}$$

3. Interchange of the Integration Limits

After interchanging the limits, the integral changes the sign (*interchange rule*):

$$\int\limits_a^b f(x)\,dx = -\int\limits_b^a f(x)\,dx. \tag{8.48}$$

4. Mean Value Theorem and Mean Value

1. Mean Value Theorem If a function $f(x)$ is continuous on the interval $[a,b]$, then there is at least one value ξ in this interval such that in the case A with $a < \xi < b$ and in the case B with $a > \xi > b$ (see 8.2.1.1, **2.**, p. 494)

$$\int\limits_a^b f(x)\,dx = (b-a)f(\xi) \tag{8.49}$$

is valid.

The geometric meaning of this theorem is that between the points a and b there exists at least one point

ξ such that the area of the figure $ABCD$ is equal to the area of the rectangle $AB'C'D$ in **Fig. 8.8**. The value

$$m = \frac{1}{b-a} \int_a^b f(x)\,dx \tag{8.50}$$

is called the *mean value* or the *arithmetic average of the function* $f(x)$ in the interval $[a, b]$.

2. Generalized Mean Value Theorem If the functions $f(x)$ and $\varphi(x)$ are continuous on the closed interval $[a, b]$, and $\varphi(x)$ does not change its sign in this interval, then there exists at least one point ξ such that

$$\int_a^b f(x)\varphi(x)\,dx = f(\xi)\int_a^b \varphi(x)\,dx \qquad (a < \xi < b) \tag{8.51}$$

is valid.

5. Estimation of the Definite Integral

The value of a definite integral lies between the values of the products of the infimum m and the supremum M of the function on the interval $[a, b]$ multiplied by the length of the interval:

$$m(b-a) \le \int_a^b f(x)\,dx \le M(b-a) \quad (a < b\,, f(x) \ge 0)\,. \tag{8.52}$$

If f is continuous, then m is the minimum and M is the maximum of the function. It is easy to recognize the geometrical interpretation of this theorem in **Fig. 8.9**.

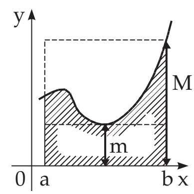


Figure 8.9

8.2.1.4 Evaluation of the Definite Integral

1. Principal Method

The principal method of calculating a definite integral is based on the fundamental theorem of integral calculus, i.e., the calculation of the indefinite integral (see 8.2.1.2, **1.**, p. 495), e.g., using **Table 21.7** p. 1065. Before substituting the limits it must be ckecked if there is an improper integral.

Nowadays computer algebra systems can be used to determine analytically the indefinite and definite integrals (see Chapter 20).

2. Transformation of Definite Integrals

In many cases, definite integrals can be calculated by appropriate transformations, with the help of the substitution method or partial integration.

■ **A:** Using the substitution method for $I = \int_0^a \sqrt{a^2 - x^2}\,dx$.

First one substitutes: $x = \varphi(t) = a\sin t,\ t = \psi(x) = \arcsin\frac{x}{a},\ \psi(0) = 0,\ \psi(a) = \frac{\pi}{2}$.

Then $I = \int_0^a \sqrt{a^2 - x^2}\,dx = \int_{\arcsin 0}^{\arcsin 1} a^2\sqrt{1-\sin^2 t}\cos t\,dt = a^2\int_0^{\frac{\pi}{2}}\cos^2 t\,dt = a^2\int_0^{\frac{\pi}{2}}\frac{1}{2}(1+\cos 2t)\,dt$.

With the further substitution $t = \varphi(z) = \frac{z}{2},\ z = \psi(t) = 2t,\ \psi(0) = 0,\ \psi(\frac{\pi}{2}) = \pi$ the value

$I = \frac{a^2}{2}\,t|_0^{\frac{\pi}{2}} + \frac{a^2}{4}\int_0^{\pi}\cos z\,dz = \frac{\pi a^2}{4} + \frac{a^2}{4}\sin z|_0^{\pi} = \frac{\pi a^2}{4}$ is obtained.

■ **B:** Method of partial integration: $\int_0^1 x\,e^x\,dx = [xe^x]_0^1 - \int_0^1 e^x\,dx = e - (e-1) = 1$.

3. Method for Calculation of More Difficult Integrals

If the determination of an indefinite integral is too difficult and complicated, or it is not possible to express it in terms of elementary functions, then there are still some further ideas to determine the value of the integral in several cases. Here the integration of functions with complex variables is mentioned (see the examples on p. 754–757) or the theorem about the differentiation of an integral with respect to a parameter (see 8.2.4, p. 512):

$$\frac{d}{dt}\int_a^b f(x,t)\,dx = \int_a^b \frac{\partial f(x,t)}{\partial t}dx. \tag{8.53}$$

■ $I = \int_0^1 \frac{x-1}{\ln x}dx$. Introducing the parameter t: $F(t) = \int_0^1 \frac{x^t-1}{\ln x}dx$; $F(0) = 0$; $F(1) = I$.
Using (8.53) for $F(t)$: $\frac{dF}{dt} = \int_0^1 \frac{\partial}{\partial t}\left[\frac{x^t-1}{\ln x}\right]dx = \int_0^1 \frac{x^t \ln x}{\ln x}dx = \int_0^1 x^t\,dx = \frac{1}{t+1}x^{t+1}\Big|_0^1 = \frac{1}{t+1}$.
Integration: $F(t) - F(0) = \int_0^t \frac{d\tau}{\tau+1} = \ln(\tau+1)|_0^t = \ln(t+1)$. Result: $I = F(1) = \ln 2$.

4. Integration by Series Expansion

If the integrand $f(x)$ can be expanded into a uniformly convergent series

$$f(x) = \varphi_1(x) + \varphi_2(x) + \cdots + \varphi_n(x) + \cdots \tag{8.54}$$

in the integration interval $[a,b]$, then the integral can be written in the form

$$\int f(x)\,dx = \int \varphi_1(x)\,dx + \int \varphi_2(x)\,dx + \cdots + \int \varphi_n(x)\,dx + \cdots. \tag{8.55}$$

In this way the definite integral can be represented as a convergent numerical series:

$$\int_a^b f(x)\,dx = \int_a^b \varphi_1(x)\,dx + \int_a^b \varphi_2(x)\,dx + \cdots + \int_a^b \varphi_n(x)\,dx + \cdots. \tag{8.56}$$

When the functions $\varphi_k(x)$ are easy to integrate, if, e.g., $f(x)$ can be expanded in a power series, which is uniformly convergent in the interval $[a,b]$, then the integral $\int_a^b f(x)\,dx$ can be calculated to arbitrary accuracy.

■ Calculate the integral $I = \int_0^{1/2} e^{-x^2}\,dx$ with an accuracy of 0.0001. The series $e^{-x^2} = 1 - \frac{x^2}{1!} + \frac{x^4}{2!} - \frac{x^6}{3!} + \frac{x^8}{4!} - \cdots$ is uniformly convergent in any finite interval according to the Abel theorem (see 7.3.3.1, p. 469), so $\int e^{-x^2}\,dx = x\left(1 - \frac{x^2}{1!\cdot 3} + \frac{x^4}{2!\cdot 5} - \frac{x^6}{3!\cdot 7} + \frac{x^8}{4!\cdot 9} - \cdots\right)$ holds. With this result it follows that $I = \int_0^{1/2} e^{-x^2}\,dx = \frac{1}{2}\left(1 - \frac{1}{2^2\cdot 1!\cdot 3} + \frac{1}{2^4\cdot 2!\cdot 5} - \frac{1}{2^6\cdot 3!\cdot 7} + \frac{1}{2^8\cdot 4!\cdot 9} - \cdots\right)$
$= \frac{1}{2}\left(1 - \frac{1}{12} + \frac{1}{160} - \frac{1}{2688} + \frac{1}{55296} - \cdots\right)$. To achieve the accuracy 0.0001 for the calculation of the integral it is enough to consider the first four terms, according to the theorem of Leibniz about alternating series (see 7.2.3.3, **1.**, p. 463):
$I \approx \frac{1}{2}(1 - 0.08333 + 0.00625 - 0.00037) = \frac{1}{2}\cdot 0.92255 = 0.46127$, $\int_0^{1/2} e^{-x^2}\,dx = 0.4613$.

5. Graphical Integration

Graphical integration is a graphical method to integrate a function $y = f(x)$ which is given by a curve AB (**Fig. 8.10**), i.e., to calculate graphically the integral $\int_a^b f(x)\,dx$, the area of the region M_0ABN:

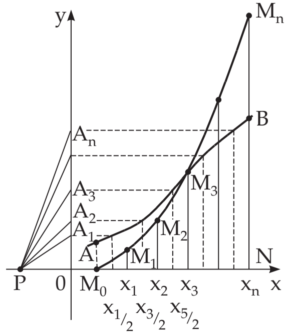


Figure 8.10

1. The interval $\overline{M_0N}$ is divided by the points $x_{1/2}, x_1, x_{3/2}, x_2, \ldots, x_{n-1}, x_{n-1/2}$ into $2n$ equal parts, where the result is more accurate if there are more points of division.
2. At the points of division $x_{1/2}, x_{3/2}, \ldots, x_{n-1/2}$ one draws vertical lines intersecting the curve. The ordinate values of the segments are denoted on the y-axis by $\overline{OA_1}, \overline{OA_2}, \ldots, \overline{OA_n}$.
3. A segment $\overline{OP}$ of arbitrary length is placed on the negative x-axis, and P is connected with the points $A_1, A_2, \ldots, A_n$.
4. Through the point M_0 a line segment is drawn parallel to PA_1 to the intersection point with the line $x = x_1$; this is the segment $\overline{M_0M_1}$. Through the point M_1 the segment $\overline{M_1M_2}$ is drawn parallel to PA_2 to the intersection with the line $x = x_2$, etc., until the last point M_n is reached with the abscissa x_n.
The integral is numerically equal to the product of the length of $\overline{OP}$ and the length of $\overline{NM_n}$:

$$\int_a^b f(x)\,dx \approx \overline{OP} \cdot \overline{NM_n}. \tag{8.57}$$

By a suitable choice of the arbitrary segment $\overline{OP}$ the extent of the graph can be influenced; the smaller the graph is wanted, the longer the segment $\overline{OP}$ should be chosen . If $\overline{OP} = 1$, then $\int_a^b f(x)\,dx = \overline{NM_n}$, and the broken line $M_0, M_1, M_2, \ldots, M_n$ represents approximately the graph of a primitive function of $f(x)$, i.e., one of the functions given by the indefinite integral $\int f(x)\,dx$.

6. Planimeter and Integraph

A *planimeter* is a tool to find the area bounded by a closed plane curve, thus also to compute a definite integral of a function $y = f(x)$ given by the curve. Special types of planimeter can evaluate not only $\int y\,dx$, but also $\int y^2\,dx$ and $\int y^3\,dx$.

An *integraph* is a device which can be used to draw the graph of a primitive function $Y = \int_a^x f(t)\,dt$ if the graph of a function $y = f(x)$ is given (see [19.30]).

7. Numerical Integration

If the integrand of a definite integral is too complicated, or the corresponding indefinite integral cannot be expressed by elementary functions, or the values of the function are known only at discrete points, e.g., from a table of values, then the so-called quadrature formulas or other methods of numerical mathematics are used (see 19.3.1, p. 963).

8.2.2 Applications of Definite Integrals

8.2.2.1 General Principle for Applications of the Definite Integral

1. The quantity A to be determined is decomposed into a large number of very small quantities, i.e., into infinitesimal quantities:

$$A = a_1 + a_2 + \cdots + a_n. \tag{8.58}$$

2. Every one of these infinitesimal quantities a_i is replaced by a quantity $\tilde{a}_i$, which differs only very slightly in value from a_i, but which can be integrated by known formulas. Here the error $\alpha_i = a_i - \tilde{a}_i$ should be an infinitesimal quantity of higher order than a_i and $\tilde{a}_i$.

3. Representation of $\tilde{a}_i$ by a variable x and a function $f(x)$ so that $\tilde{a}_i$ has the form $f(x_i)\,\Delta x_i$.

4. Evaluation of the desired quantity as the limit of the sum

$$A = \lim_{n\to\infty} \sum_{i=1}^{n} \tilde{a}_i = \lim_{n\to\infty} \sum_{i=1}^{n} f(x_i)\Delta x_i = \int_a^b f(x)\,dx, \tag{8.59}$$

where $\Delta x_i \ge 0$ holds for every i. The lower and upper limit for x is denoted by a and b.

■ Evaluating the volume V of a pyramid with base area S and height H **(Fig. 8.11a-c)**:

a) Decomposition of the required volume V by plane sections into frustums **(Fig. 8.11a)**: $V = v_1 + v_2 + \cdots + v_n$.

b) Replacing every frustum by a prism, whose volume is $\tilde{v}_i$, with the same height and with a base area of the top base of the frustum **(Fig. 8.11b)**. The difference of their volumes is an infinitesimal quantity of higher order than v_i.

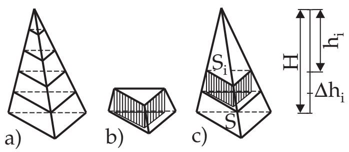


Figure 8.11

c) Representation of the volume $\tilde{v}_i$ in the form $\tilde{v}_i = S_i\,\Delta h_i$, where h_i **(Fig. 8.11c)** is the distance of the top surface from the vertex of the pyramid. Since $S_i : S = h_i^2 : H^2$ one can write: $\tilde{v}_i = \dfrac{Sh_i^2}{H^2}\,\Delta h_i$.

d) Calculation of the limit of the sum

$$V = \lim_{n\to\infty} \sum_{i=1}^{n} \tilde{v}_i = \lim_{n\to\infty} \sum_{i=1}^{n} \frac{Sh_i^2}{H^2}\,\Delta h_i = \int_0^H \frac{Sh^2}{H^2}\,dh = \frac{SH}{3}.$$

8.2.2.2 Applications in Geometry

1. Area of Planar Figures

1. Area of a Curvilinear Trapezoid Between B and C (Fig. 8.12a) if the curve is given by an equation in explicit form ($y = f(x)$ and $a \le x \le b$) or in parametric form ($x = x(t),\ y = y(t),\ t_1 \le t \le t_2$):

$$S_{\mathrm{ABCD}} = \int_a^b f(x)\,dx = \int_{t_1}^{t_2} y(t)x'(t)\,dt \ \ (f(x) \ge 0; x(t_1) = a; x(t_2) = b; y(t) \ge 0)\,. \tag{8.60a}$$

2. Area of a Curvilinear Trapezoid Between G and H (Fig. 8.12b) if the curve is given by an equation in explicit form ($x = g(y)$ and $\alpha \le y \le \beta$) or in parametric form ($x = x(t),\ y = y(t),\ t_1 \le t \le t_2$):

$$S_{\mathrm{EFGH}} = \int_\alpha^\beta g(y)\,dy = \int_{t_1}^{t_2} x(t)y'(t)\,dt \ \ (g(y) \ge 0; y(t_1) = \alpha; y(t_2) = \beta; x(t) \ge 0)\,. \tag{8.60b}$$

3. Area of a Curvilinear Sector (Fig. 8.12c), bounded by a curve between K and L, which is given by an equation in polar coordinates ($\rho = \rho(\varphi),\ \varphi_1 \le \varphi \le \varphi_2$):

$$S_{\mathrm{OKL}} = \frac{1}{2}\int_{\varphi_1}^{\varphi_2} \rho^2\,d\varphi. \tag{8.60c}$$

Areas of more complicated figures can be calculated by partition of the area into simple parts, or by line integrals (see 8.3, p. 515) or by double integrals (see 8.4.1, p. 524).

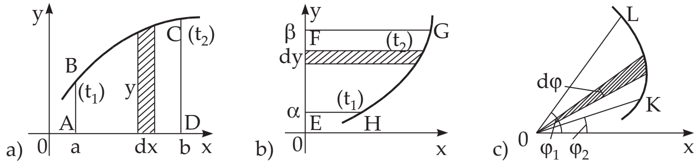


Figure 8.12

2. Arc-length of Plane Curves

1. Arc-length of a Curve Between Two Points (I) A and B, given in explicit form ($y = f(x)$ or $x = g(y)$) or in parametric form ($x = x(t)$, $y = y(t)$) **(Fig. 8.13a)** can be calculated by the integrals:

$$L_{\stackrel{\frown}{AB}} = \int\limits_a^b \sqrt{1 + [f'(x)]^2}\,dx = \int\limits_\alpha^\beta \sqrt{[g'(y)]^2 + 1}\,dy = \int\limits_{t_1}^{t_2} \sqrt{[x'(t)]^2 + [y'(t)]^2}\,dt. \tag{8.61a}$$

With the differential of the arc-length dl one gets

$$L = \int dl \qquad \text{with} \qquad dl^2 = dx^2 + dy^2. \tag{8.61b}$$

■ The perimeter of the ellipse with the help of (8.61a): With the substitutions $x = x(t) = a\sin t,\ y = y(t) = b\cos t$ it follows that $L_{\stackrel{\frown}{AB}} = \int_{t_1}^{t_2} \sqrt{a^2 - (a^2 - b^2)\sin^2 t}\,dt = a\int_{t_1}^{t_2} \sqrt{1 - e^2\sin^2 t}\,dt$, where $e = \sqrt{a^2 - b^2}/a$ is the numerical eccentricity of the ellipse.
Since $x = 0,\ y = b$ and $x = a,\ y = 0$ the limits of the integral in the first quadrant are $t_1 = 0$ and $t_2 = \pi/2$ so, for the perimeter of the ellipse holds $L_{\stackrel{\frown}{AB}} = 4a\int_0^{\pi/2} \sqrt{1 - e^2\sin^2 t}\,dt = a\,E(k, \frac{\pi}{2})$ with $k = e$. The value of the integral $E(k, \frac{\pi}{2})$ is given in **Table 21.9** (see example in 8.1.4.3, p. 491).

2. Arc-length of a Curve Between Two Points (II) C and D, given in polar coordinates ($\rho = \rho(\varphi)$) **(Fig. 8.13b)**:

$$L_{\stackrel{\frown}{CD}} = \int\limits_{\varphi_1}^{\varphi_2} \sqrt{\rho^2 + \left(\frac{d\rho}{d\varphi}\right)^2}\,d\varphi. \tag{8.61c}$$

With the differential of the arc-length dl one gets

$$L = \int dl \quad \text{with} \quad dl^2 = \rho^2 d\varphi^2 + d\rho^2. \tag{8.61d}$$

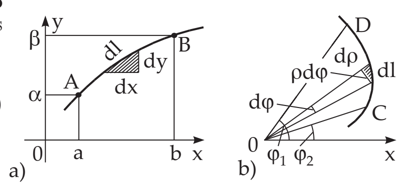


Figure 8.13

3. Surface Area of a Body of Revolution (see also First Guldin Rule, p. 506)

1. The area of the surface of a body given by rotating the graph of the function $y = f(x) \geq 0$ around the x-axis **(Fig. 8.14a)** is:

$$S = 2\pi\int\limits_a^b y\,dl = 2\pi\int\limits_a^b y(x)\sqrt{1 + \left(\frac{dy}{dx}\right)^2}\,dx. \tag{8.62a}$$

2. The area of the surface of a body given by rotating $x = f(y) \geq 0$ around the y-axis **(Fig. 8.14b)** is:

$$S = 2\pi \int_{\alpha}^{\beta} x\,dl = 2\pi \int_{\alpha}^{\beta} x(y)\sqrt{\left(1+\frac{dx}{dy}\right)^2}\,dy. \tag{8.62b}$$

3. To calculate the area of more complicated surfaces see the application of double integrals in 8.4.1.3, p. 527 and the application of surface integrals of the first kind, 8.5.1.3, p. 535. General formulas for the calculation of surface areas with double integrals are given in **Table 8.9** (Applications of double integrals), p. 528.

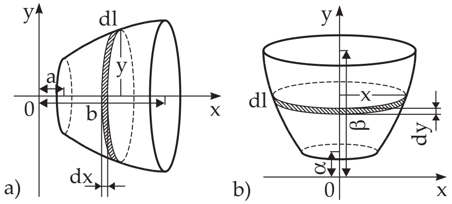


Figure 8.14

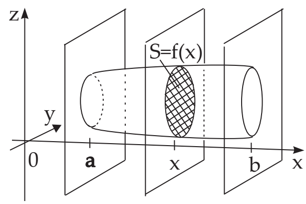


Figure 8.15

4. Volume (see also Second Guldin Rule, p. 506)

1. The volume of a rotationally symmetric body given by rotation around the x-axis **(Fig. 8.14a)** is:

$$V = \pi \int_{a}^{b} y^2\,dx. \tag{8.63a}$$

2. The volume of a rotationally symmetric body given by rotation around the y-axis **(Fig. 8.14b)** is:

$$V = \pi \int_{a}^{b} x^2\,dy. \tag{8.63b}$$

3. The volume of a body, whose section perpendicular to the x-axis **(Fig. 8.15)** has an area given by the function $S = f(x)$, is:

$$V = \int_{a}^{b} f(x)\,dx. \tag{8.64a}$$

■ Calculation of the volume of an ellipsoid of revolution centered at the origin. Since the ellipsoid of revolution $x^2/a^2 + y^2/b^2 + z^2/c^2 = 1$ (see (3.412), 3.5.3.13,**2.**, p. 224 and **Fig.8.16**) with $a = c$ arises by rotating the ellipse $y^2/b^2 + z^2/c^2 = 1$ about the y-axis, the area of the circle shaped cross-sections being parallel to the x, z-plane is $S = f(y) = \pi z^2 = \pi c^2(1 - y^2/b^2)$, so by integration $V = 2\pi c^2 \int_0^b (1 - y^2/b^2)dy = (4/3)\pi b c^2$.

4. Cavalieri Principle If in the interval $[a, b]$ there exists in addition to a cross-sectional area function $S = f(x)$ a second cross-sectional area function $\overline{S} = \overline{f}(x)$ which has the same value for every x as $f(x)$, then the volumes V according to (8.64a) and $\overline{V}$ according to (8.64b) are equal:

$$\overline{V} = \int_{a}^{b} \overline{f}(x) = V\,. \tag{8.64b}$$

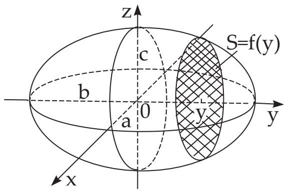


Figure 8.16

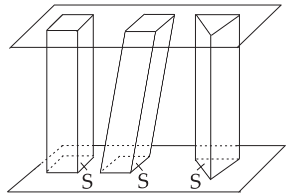


Figure 8.17

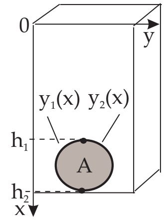


Figure 8.18

The original Cavalieri's principle tells: If two bodies are bounded below and above by two parallel planes and in these planes and in every plane being parallel to these ones their cross-sectional areas are equal to each other, then the volumes of these bodies are also equal (**Fig.8.17**).

5. Tables General formulas to calculate volumes with multiple integrals are given in **Table 8.9** (Applications of double integrals, see p. 528) and **Table 8.11** (Applications of triple integrals, see p. 533).

8.2.2.3 Applications in Mechanics and Physics

1. Distance Traveled by a Point

The distance traveled by a moving point during the time from t_0 until T with a time-dependent velocity $v = f(t)$ is

$$s = \int_{t_0}^{T} v\,dt. \tag{8.65}$$

2. Work

The work in moving a body in a force field depends of the direction of motion. Here it is supposed that the direction of the field and the direction of the movement are constant and coincide along the x-axis. If the magnitude of the force $\vec{F}$ is changing, i.e., $|\vec{F}| = f(x)$, then the work W necessary to move the body from the point $x = a$ to the point $x = b$ along the x-axis is

$$W = \int_{a}^{b} f(x)\,dx. \tag{8.66}$$

In the general case, when the direction of the force field and the direction of the movement are not coincident, the work is calculated as a line integral (see (8.130), p. 522) of the scalar product of the force and the variation of the position vector at every point of $\vec{\mathbf{r}}$ along the given path.

3. Gravitational and Lateral Pressure

In a fluid at rest in the gravitational field of the Earth with a density ϱ and gravitational acceleration g (see **Table 21.2**, p. 1053) the *gravitational pressure* p and the *lateral pressure* p_s are distinguished. The gravitational pressure p at a depth x under the surface of the fluid (see **Fig.8.18**) is

$$p = \varrho\, g\, x\,. \tag{8.67a}$$

The lateral pressure p_s acting e.g. on a cover plate (surface area A) of a side opening of a container (see **Fig.8.18**), is caused by the pressure force F acting in all directions. The differential pressure force dF acting perpendicular on a side surface-element dA in depth x under the surface of the fluid is

$$d\,F = \varrho\, g\, x\, d\,A = \varrho\, g\, x\, y(x)\, d\,x\,. \tag{8.67b}$$

Integration and division by A yields

$$p_s = \frac{\varrho\, g}{A} \int_{h_1}^{h_2} x(y_2(x) - y_1(x))\, dx = \varrho\, g\, x_C\,. \tag{8.67c}$$

Functions $y_1(x)$ and $y_2(x)$ describe the left and right boundaries of the cover plate and x_C is the x-coordinate of its center of mass (see center of gravity of planar figures paragraph **5.**, p. 505).

Remark: The center of mass of the cover plate usually does not coincide with the point of impact of pressure force F since the lateral pressure force is proportional to x.

4. Moments of Inertia

1. Moment of Inertia of an Arc The moment of inertia of a homogeneous curve segment $y = f(x)$ with constant density ϱ in the interval $[a, b]$ with respect to the y-axis **(Fig. 8.19a)** is:

$$I_y = \varrho \int_a^b x^2\, dl = \varrho \int_a^b x^2 \sqrt{1 + (y')^2}\, dx. \tag{8.68}$$

If the density is a function $\varrho(x)$, then its analytic expression is in the integrand.

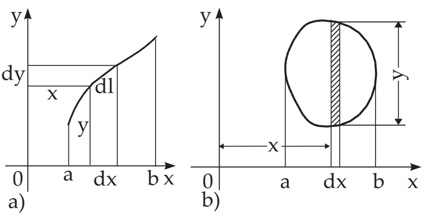


Figure 8.19

2. Moment of Inertia of a Planar Figure The moment of inertia of a planar figure with a homogeneous density ϱ with respect to the y-axis, where y is the length of the cut parallel to the y-axis **(Fig. 8.19b)**, is:

$$I_y = \varrho \int_a^b x^2 y\, dx. \tag{8.69}$$

(See also **Table 8.9**, (Applications of the Double Integral), p. 528.) If the density is position dependent, then its analytic expression must be in the integrand.

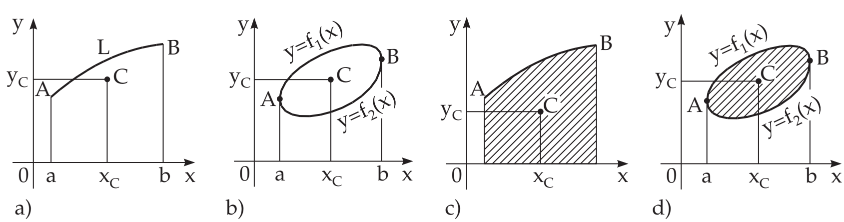


Figure 8.20

5. Center of Gravity, Guldin Rules

1. Center of Gravity of an Arc Segment The center of gravity C of an arc segment of a homogeneous plane curve $y = f(x)$ in the interval $[a, b]$ with a length L **(Fig. 8.20a)** considering (8.61a), p. 502, has the coordinates:

$$x_C = \frac{\int_a^b x\sqrt{1 + y'^2}\, dx}{L}, \qquad y_C = \frac{\int_a^b y\sqrt{1 + y'^2}\, dx}{L}. \tag{8.70}$$

2. Center of Gravity of a Closed Curve The center of gravity C of a closed curve $y = f(x)$ **(Fig. 8.20b)** with the equations $y_1 = f_1(x)$ for the upper part and $y_2 = f_2(x)$ for the lower part, and with a length L has the coordinates:

$$x_C = \frac{\int_a^b x\left(\sqrt{1+(y_1')^2} + \sqrt{1+(y_2')^2}\right) dx}{L}, \qquad y_C = \frac{\int_a^b \left(y_1\sqrt{1+(y_1')^2} + y_2\sqrt{1+(y_2')^2}\right) dx}{L}. \tag{8.71}$$

3. First Guldin Rule Suppose a plane curve segment is rotated around an axis which lies in the plane of the curve and does not intersect the curve. Choosing it as the x-axis the surface area S_{rot} of the body generated by the rotated curve segment is the product of the perimeter of the circle drawn by the center of gravity at a distance r_C from the axis of rotation, i.e., $2\pi r_C$, and the length of the curve segment L:

$$S_{\text{rot}} = L \cdot 2\pi r_C. \tag{8.72}$$

4. Center of Gravity of a Trapezoid The center of gravity C of a homogeneous trapezoid bounded above by a curve segment between the points of the curve A and B **(Fig. 8.20c)**, with an area S of the trapezoid, and with the equation $y = f(x)$ of the curve segment AB, has the coordinates:

$$x_C = \frac{\int_a^b x\,y\,dx}{S}, \qquad y_C = \frac{\frac{1}{2}\int_a^b y^2\,dx}{S}. \tag{8.73}$$

5. Center of Gravity of an Arbitrary Planar Figure The center of gravity C of an arbitrary planar figure **(Fig. 8.20d)** with area S, bounded above and below by the curve segments with the equations $y_1 = f_1(x)$ and $y_2 = f_2(x)$, has the coordinates

$$x_C = \frac{\int_a^b x(y_1 - y_2)\,dx}{S}, \qquad y_C = \frac{\frac{1}{2}\int_a^b (y_1^2 - y_2^2)\,dx}{S}. \tag{8.74}$$

Formulas to calculate the center of gravity with multiple integrals are given in **Table 8.9** (Applications of double integrals, p. 528) and in **Table 8.11** (Applications of triple integrals, p. 533).

6. Second Guldin Rule Suppose a plane figure is rotated around an axis which is in the plane of the figure and does not intersect it. Choosing it as the x-axis the volume V of the body generated by the rotated figure is equal to the product of the perimeter of the circle drawn by the center of gravity under the rotation, i.e., $2\pi r_C$, and the area of the figure S:

$$V_{\text{rot}} = S \cdot 2\pi r_C. \tag{8.75}$$

8.2.3 Improper Integrals, Stieltjes and Lebesgue Integrals

8.2.3.1 Generalization of the Notion of the Integral

The notion of the definite integral (see 8.2.1.1, p. 493), as a Riemann integral (see 8.2.1.1, **2.**, p. 494), was introduced under the assumptions that the function $f(x)$ is bounded, and the interval $[a, b]$ is closed and finite. Both assumptions can be relaxed in the generalizations of the Riemann integral. Some of them are mentioned in the following section.

1. Improper Integrals

These are the generalization of the integral to unbounded functions and to unbounded intervals. In the next paragraphs *integrals with infinite integration limits* and *integrals with unbounded integrands* are discussed.

2. Stieltjes Integral for Functions of One Variable

Considering two finite functions $f(x)$ and $g(x)$ defined on the finite interval $[a, b]$ and decomposing the interval into subintervals, just as with the Riemann integral, but instead of the Riemann sum (8.38)

the following sum is constructed:

$$\sum_{i=1}^{n} f(\xi_i)[g(x_i) - g(x_{i-1})]. \tag{8.76}$$

If the limit of (8.76) exists, when the length of the subintervals tends to zero, and it is independent of the choice of the points x_i and ξ_i, then this limit is called a *definite Stieltjes integral* (see also [8.8]).

■ For $g(x) = x$ the Stieltjes integral becomes the Riemann integral.

3. Lebesgue Integral

Another generalization of the integral notion is connected with measure theory (see 12.9, p. 693), where the measure of a set, measure spaces, and measurable functions are introduced. In functional analysis the Lebesgue integral is defined (see 12.9.3.2, p. 696) based on these notions (see [8.6]). The generalization with comparison to the Riemann integral is, e.g., the domain of integration can be a rather general subset of $\mathbb{R}^n$ and it is partitioned into measurable subsets.

There are different notations for the generalizations of the integrals (see [8.8]).

8.2.3.2 Integrals with Infinite Integration Limits

1. Definitions

a) If the integration domain is the closed half-axis $[a, +\infty)$, and if the integrand is defined there, then the integral is by definition

$$\int\limits_a^{+\infty} f(x)\,dx = \lim_{B\to\infty} \int\limits_a^{B} f(x)\,dx. \tag{8.77}$$

If the limit exists, then the integral is called a *convergent improper integral*. If the limit does not exist, then the improper integral (8.77) is divergent.

b) If the domain of a function is the closed half-axis $(-\infty, b]$ or the whole real axis $(-\infty, +\infty)$, then one defines analogously the improper integrals

$$\int\limits_{-\infty}^{b} f(x)\,dx = \lim_{A\to-\infty} \int\limits_A^{b} f(x)\,dx, \tag{8.78a}$$

$$\int\limits_{-\infty}^{+\infty} f(x)\,dx = \lim_{\substack{A\to-\infty\\ B\to\infty}} \int\limits_A^{B} f(x)\,dx. \tag{8.78b}$$

c) At the limits of (8.78b) the numbers A and B tend to infinity independently of each other. If the limit (8.78b) does not exist, but the limit

$$\lim_{A\to\infty} \int\limits_{-A}^{+A} f(x)\,dx, \tag{8.78c}$$

exists, then this limit (8.78c) is called the *principal value of the improper integral*, or *Cauchy's principal value*.

Remark: An obviously necessary but not sufficient condition for the convergence of the integral (8.77) is $\lim\limits_{x\to\infty} f(x) = 0$.

2. Geometrical Meaning of Integrals with Infinite Limits

The integrals (8.77), (8.78a) and (8.78b) give the area of the figures represented in **Fig. 8.21**.

■ **A:** $\displaystyle\int_1^{\infty} \frac{dx}{x} = \lim_{B\to\infty} \int_1^{B} \frac{dx}{x} = \lim_{B\to\infty} \ln B = \infty$ (divergent).

■ **B:** $\displaystyle\int_2^{\infty} \frac{dx}{x^2} = \lim_{B\to\infty} \int_2^{B} \frac{dx}{x^2} = \lim_{B\to\infty} \left(\frac{1}{2} - \frac{1}{B}\right) = \frac{1}{2}$ (convergent).

■ **C:** $\int_{-\infty}^{+\infty} \frac{dx}{1+x^2} = \lim_{\substack{A\to-\infty\\B\to+\infty}} \int_A^B \frac{dx}{1+x^2} = \lim_{\substack{A\to-\infty\\B\to+\infty}} [\arctan B - \arctan A] = \frac{\pi}{2} - \left(-\frac{\pi}{2}\right) = \pi$ (convergent).

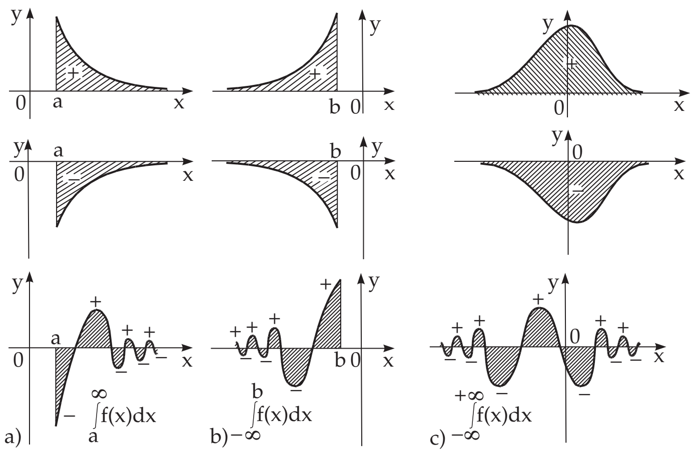


Figure 8.21

3. Sufficient Criteria for Convergence

If the direct calculation of the limits (8.77), (8.78a) and (8.78b) is complicated, or if only the convergence or divergence of an improper integral is the question, then one of the following sufficient criteria can be used. Here, only the integral (8.77) is considered. The integral (8.78a) can be transformed into (8.77) by substitution of x by $-x$:

$$\int_{-\infty}^{a} f(x)\,dx = \int_{-a}^{+\infty} f(-x)\,dx. \tag{8.79}$$

The integral (8.78b) can be decomposed into the sum of two integrals of type (8.77) and (8.78a):

$$\int_{-\infty}^{+\infty} f(x)\,dx = \int_{-\infty}^{c} f(x)\,dx + \int_{c}^{+\infty} f(x)\,dx, \tag{8.80}$$

where c is an arbitrary number.

Criterion 1: If f is integrable in any finite interval of $[a, \infty)$, and if the integral

$$\int_{a}^{+\infty} |\, f(x)\,|\; dx \tag{8.81}$$

is convergent, then also the integral (8.77) is convergent. In this case the integral (8.77) is called *absolutely convergent*, and the function $f(x)$ is *absolute integrable* on the half-axis $[a, +\infty)$.

Criterion 2: If for the functions $f(x)$ and $\varphi(x)$

$$f(x) > 0, \;\; \varphi(x) > 0 \quad \text{and} \quad f(x) \le \varphi(x) \;\text{ for }\; a \le x < +\infty \tag{8.82a}$$

hold, then from the convergence of the integral

$$\int_a^{+\infty} \varphi(x)\,dx \qquad (8.82b) \qquad \text{the convergence of the integral} \quad \int_a^{+\infty} f(x)\,dx \qquad (8.82c)$$

follows, and conversely, from the divergence of the integral (8.82c) the divergence of the integral (8.82b) follows.

Criterion 3: Substituting (8.83a) and considering that for $a > 0,\ \alpha > 1$ the integral (8.83b)

$$\varphi(x) = \frac{1}{x^\alpha}, \qquad (8.83a) \qquad \int_a^{+\infty} \frac{dx}{x^\alpha} = \frac{1}{(\alpha-1)a^{\alpha-1}} \quad (a > 0,\ \alpha > 1) \qquad (8.83b)$$

is convergent and has the value of the right-hand side, and the integral of the left-hand side is divergent for $\alpha \le 1$, then one can deduce a further convergence criterion from the second one:

If $f(x)$ in $a \le x < \infty$ is a positive function, and there exists a number $\alpha > 1$ such that for all x large enough

$$f(x)\,x^\alpha < k < \infty \quad (k > 0,\ \text{const}) \qquad (8.83c)$$

holds, then the integral (8.77) is convergent; if $f(x)$ is positive and there exists a number $\alpha \le 1$ such that

$$f(x)\,x^\alpha > c > 0 \quad (c > 0,\ \text{const}) \qquad (8.83d)$$

holds from a certain point x, then the integral (8.77) is divergent.

■ $\int_0^{+\infty} \frac{x^{3/2}\,dx}{1+x^2}$. Substituting $\alpha = \frac{1}{2}$, gives $\frac{x^{3/2}}{1+x^2}\,x^{1/2} = \frac{x^2}{1+x^2} \to 1$. The integral is divergent.

4. Relations Between Improper Integrals and Infinite Series

If $x_1, x_2, \ldots, x_n, \ldots$ is an arbitrary, unlimited increasing infinite sequence, i.e., if

$$a < x_1 < x_2 < \cdots < x_n < \cdots \quad \text{with} \quad \lim_{n\to+\infty} x_n = \infty, \qquad (8.84a)$$

and if the function $f(x)$ is positive for $a \le x < \infty$, then the problem of convergence of the integral (8.77) can be reduced to the problem of convergence of the series

$$\int_a^{x_1} f(x)\,dx + \int_{x_1}^{x_2} f(x)\,dx + \cdots + \int_{x_{n-1}}^{x_n} f(x)\,dx + \cdots. \qquad (8.84b)$$

If the series (8.84b) is convergent, then the integral (8.77) is also convergent, and it is equal to the sum of the series (8.84b). If the series (8.84b) is divergent, then the integral (8.77) is also divergent. So the convergence criteria for series can be used for improper integrals, and conversely, in the integral criterion for series (see 7.2.2.4, p. 462) one can use the improper integrals to investigate the convergence of infinite series.

8.2.3.3 Integrals with Unbounded Integrand

1. Definitions

1. Right Open Interval For a function $f(x)$, which has a domain $[a, b)$ open on the right, and at the point b it has the improper limit $\lim_{x\to b-0} f(x) = \infty$, the definition of the improper integral is the following:

$$\int_a^b f(x)\,dx = \lim_{\varepsilon\to+0} \int_a^{b-\varepsilon} f(x)\,dx. \qquad (8.85)$$

If this limit exists and is finite, then the improper integral (8.85) exists, and is called a *convergent improper integral*. If the limit does not exist or it is not finite, then the integral is called a *divergent improper integral*.

2. Left Open Interval For a function $f(x)$, which has a domain open on the left $(a, b]$, and at the point a it has the limit $\lim\limits_{x\to a+0} f(x) = \infty$, the definition of the improper integral analogously to (8.85) is

$$\int\limits_a^b f(x)\,dx = \lim_{\varepsilon\to+0} \int\limits_{a+\varepsilon}^b f(x)\,dx. \tag{8.86}$$

3. Two Half-Open Continuous Intervals For a function $f(x)$, which is defined on the interval $[a, b]$ except at an interior point c with $a < c < b$, i.e., for a function $f(x)$ defined on the half-open intervals $[a, c)$ and $(c, b]$, or is defined on the interval $[a, b]$, but at the interior point c it has an infinite limit at least from one side $\lim\limits_{x\to c+0} f(x) = \infty$ or $\lim\limits_{x\to c-0} f(x) = \infty$, the definition of the improper integral is

$$\int\limits_a^b f(x)\,dx = \lim_{\varepsilon\to+0} \int\limits_a^{c-\varepsilon} f(x)\,dx + \lim_{\delta\to+0} \int\limits_{c+\delta}^b f(x)\,dx. \tag{8.87a}$$

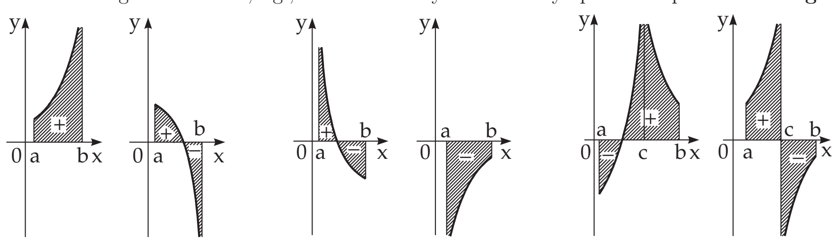
Here the numbers ε and δ tend to zero independently of each other. If the limit (8.87a) does not exist, but the limit

$$\lim_{\varepsilon\to+0} \left\{ \int\limits_a^{c-\varepsilon} f(x)\,dx + \int\limits_{c+\varepsilon}^b f(x)\,dx \right\} \tag{8.87b}$$

does, then the limit (8.87b) is called the *principal value of the improper integral* or *Cauchy's principal value*.

2. Geometrical Meaning

The geometrical meaning of the integrals of unbounded functions (8.85), (8.86), and (8.87a) is to find the area of the figures bounded, e.g., from one side by a vertical asymptote as represented in **Fig.8.22**.



a) s. (8.85) b) s. (8.86) c) s. (8.87a)

Figure 8.22

■ **A:** $\displaystyle\int_0^b \frac{dx}{\sqrt{x}}$: Case (8.86), singular point at $x = 0$.

$$\int_0^b \frac{dx}{\sqrt{x}} = \lim_{\varepsilon\to+0} \int_\varepsilon^b \frac{dx}{\sqrt{x}} = \lim_{\varepsilon\to+0} (2\sqrt{b} - 2\sqrt{\varepsilon}) = 2\sqrt{b} \quad \text{(convergent)}.$$

■ **B:** $\displaystyle\int_0^{\pi/2} \tan x\,dx$: Case (8.85), singular point at $x = \dfrac{\pi}{2}$.

$$\int_0^{\pi/2} \tan x\,dx = \lim_{\varepsilon\to+0} \int_0^{\pi/2-\varepsilon} \tan x\,dx = \lim_{\varepsilon\to+0} \left[\ln\cos 0 - \ln\cos\left(\frac{\pi}{2} - \varepsilon\right)\right] = \infty \quad \text{(divergent)}.$$

■ **C:** $\int_{-1}^{8} \frac{dx}{\sqrt[3]{x}}$: Case (8.87a), singular point at $x = 0$.

$$\int_{-1}^{8} \frac{dx}{\sqrt[3]{x}} = \lim_{\varepsilon\to+0} \int_{-1}^{-\varepsilon} \frac{dx}{\sqrt[3]{x}} + \lim_{\delta\to+0} \int_{\delta}^{8} \frac{dx}{\sqrt[3]{x}} = \lim_{\varepsilon\to+0} \frac{3}{2}(\varepsilon^{2/3} - 1) + \lim_{\delta\to+0} \frac{3}{2}(4 - \delta^{2/3}) = \frac{9}{2} \text{ (convergent)}.$$

■ **D:** $\int_{-2}^{2} \frac{2x\,dx}{x^2 - 1}$: Case (8.87a), singular point at $x = \pm 1$.

$$\int_{-2}^{2} \frac{2x\,dx}{x^2-1} = \lim_{\varepsilon\to+0} \int_{-2}^{-1-\varepsilon} + \lim_{\substack{\delta\to+0\\ \nu\to+0}} \int_{-1+\delta}^{1-\nu} + \lim_{\gamma\to+0} \int_{1+\gamma}^{2}$$

$$= \lim_{\varepsilon\to+0} \ln|x^2-1| \,|_{-2}^{-1-\varepsilon} + \cdots = \lim_{\varepsilon\to+0} [\ln|1 + 2\varepsilon + \varepsilon^2 - 1| - \ln 3] + \cdots = \infty \quad \text{(divergent)}.$$

3. The Application of the Fundamental Theorem of Integral Calculus

1. Warning The calculation of improper integrals of type (8.87a) with the incorrect use of the formula

$$\int_{a}^{b} f(x)\,dx = F(x)\,|_a^b \quad \text{with} \quad F'(x) = f(x) \tag{8.88}$$

(see 8.2.1.2,**1.**, p. 495) usually results in mistakes if the singular points in the interval $[a, b]$ are not taken into consideration.

■ **E:** Using formally the fundamental theorem one gets for the example **D**

$$\int_{-2}^{2} \frac{2x\,dx}{x^2-1} = \ln|x^2-1|\,|_{-2}^{2} = \ln 3 - \ln 3 = 0,$$

though this integral is divergent.

2. General Rule The fundamental theorem of integral calculus can be used for (8.87a) only if the primitive function of $f(x)$ can be defined to be continuous at the singular point.

■ **F:** In example **D** the function $\ln|x^2 - 1|$ is discontinuous at $x = \pm 1$, so the conditions are not fulfilled. Considering example **C**, the function $y = \frac{3}{2}x^{2/3}$ is such a primitive function of $\frac{1}{\sqrt[3]{x}}$ on the intervals $[a, 0)$ and $(0, b]$ which can be defined continuously at $x = 0$, so the fundamental theorem can be used in example **C**:

$$\int_{-1}^{8} \frac{dx}{\sqrt[3]{x}} = \frac{3}{2}x^{2/3}\,|_{-1}^{8} = \frac{3}{2}(8^{2/3} - (-1)^{2/3}) = \frac{9}{2}.$$

4. Sufficient Conditions for the Convergence of an Improper Integral with Unbounded Integrand $\lim_{x\to b-0} f(x) = \infty$

1. If the improper integral $\int_a^b |f(x)|\,dx$ converges, then the improper integral $\int_a^b f(x)\,dx$ also converges. In this case it is called an *absolutely convergent integral* and the function $f(x)$ is an *absolutely integrable function* on the considered interval.

2. If the function $f(x)$ is positive in the interval $[a, b)$, and there is a number $\alpha < 1$ such that for the values of all x close enough to b

$$f(x)\,(b-x)^{\alpha} < \infty \tag{8.89a}$$

holds, then the integral (8.87a) is convergent. But, if the function $f(x)$ is positive in the interval $[a, b)$, and there is a number $\alpha > 1$ such that for the values of x close enough to b

$$f(x)\,(b-x)^{\alpha} > c > 0 \quad (c \text{ const}) \tag{8.89b}$$

holds, then the integral (8.87a) is divergent.

8.2.4 Parametric Integrals

8.2.4.1 Definition of Parametric Integrals

The definite integral

$$\int_a^b f(x,y)\,dx = F(y) \tag{8.90}$$

is a function of the variable y considered here as a parameter. In several cases the function $F(y)$ is no longer elementary, even if $f(x,y)$ is an elementary function of x and y. The integral (8.90) can be an ordinary integral, or an convergent improper integral with infinite limits or unbounded integrand $f(x,y)$.

For theoretical discussions about the convergence of improper integrals depending on a parameter see, e.g.,[8.3].

■ **Gamma Function or Euler Integral of the Second Kind** (see 8.2.5, **6.**, p. 514):

$$\Gamma(y) = \int_0^\infty x^{y-1}e^{-x}\,dx \qquad \text{(convergent for } y>0\text{).} \tag{8.91}$$

8.2.4.2 Differentiation Under the Symbol of Integration

1. Theorem If the function (8.90) is defined in the interval $c \le y \le e$, and the function $f(x,y)$ is continuous on the rectangle $a \le x \le b,\ c \le y \le e$ and it has here a continuous partial derivative with respect to y, then for arbitrary y in the interval $[c,e]$ holds

$$\frac{d}{dy}\int_a^b f(x,y)\,dx = \int_a^b \frac{\partial f(x,y)}{\partial y}\,dx. \tag{8.92}$$

This is called *differentiation under the symbol of integration.*

■ For arbitrary $y>0$: $\displaystyle \frac{d}{dy}\int_0^1 \arctan\frac{x}{y}\,dx = \int_0^1 \frac{\partial}{\partial y}\left(\arctan\frac{x}{y}\right)dx = -\int_0^1 \frac{x\,dx}{x^2+y^2} = \frac{1}{2}\ln\frac{y^2}{1+y^2}\,.$

Checking: $\displaystyle \int_0^1 \arctan\frac{x}{y}\,dx = \arctan\frac{1}{y} + \frac{1}{2}y\ln\frac{y^2}{1+y^2}\,;\ \frac{d}{dy}\left(\arctan\frac{1}{y} + \frac{1}{2}y\ln\frac{y^2}{1+y^2}\right) = \frac{1}{2}\ln\frac{y^2}{1+y^2}\,.$

For $y=0$ the condition of continuity for $f(x,y)$ is not fulfilled, and there exists no derivative.

2. Generalization for Limits of Integration Depending on Parameters The formula (8.92) can be generalized, if with the same assumptions as have been made for (8.92) the functions $\alpha(y)$ and $\beta(y)$ are defined in the interval $[c,e]$, they are continuous and differentiable there, and the curves $x=\alpha(y),\ x=\beta(y)$ do not leave the rectangle $a \le x \le b,\ c \le y \le e$:

$$\frac{d}{dy}\int_{\alpha(y)}^{\beta(y)} f(x,y)\,dx = \int_{\alpha(y)}^{\beta(y)} \frac{\partial f(x,y)}{\partial y}\,dx + \beta'(y)\,f\left(\beta(y),y\right) - \alpha'(y)\,f\left(\alpha(y)\,,y\right). \tag{8.93}$$

8.2.4.3 Integration Under the Symbol of Integration

If the function $f(x,y)$ is continuous on the rectangle $a \le x \le b,\ c \le y \le e$, then the function (8.90) is defined in the interval $[c,e]$, and

$$\int_c^e \left[\int_a^b f(x,y)\,dx\right] dy = \int_a^b \left[\int_c^e f(x,y)\,dy\right] dx \tag{8.94}$$

is valid. This case of commutability of the order of integration is called *integration under the symbol of integration.*

■ **A:** Integration of the function $f(x, y) = x^y$ on the rectangle $0 \le x \le 1$, $a \le y \le b$. The function x^y is discontinuous at $x = 0$, $y = 0$, for $a > 0$ it is continuous. So one can change the order of integration: $\int_a^b \left[\int_0^1 x^y\,dx\right] dy = \int_0^1 \left[\int_a^b x^y\,dy\right] dx$. On the left-hand side one gets $\int_a^b \frac{dy}{1+y} = \ln\frac{1+b}{1+a}$, on the right-hand side $\int_0^1 \frac{x^b - x^a}{\ln x}\,dx$. The corresponding indefinite integral cannot be expressed by elementary functions. Anyway, the definite integral is known, so now follows:

$$\int_0^1 \frac{x^b - x^a}{\ln x}\,dx = \ln\frac{1+b}{1+a} \qquad (0 < a < b).$$

■ **B:** Integration of the function $f(x, y) = \frac{y^2 - x^2}{(x^2 + y^2)^2}$ over the rectangle $0 \le x \le 1$, $0 \le y \le 1$. The function is discontinuous at the point $(0, 0)$, so the formula (8.94) cannot be used. Checking it yields:

$$\int_0^1 \frac{y^2 - x^2}{(x^2+y^2)^2}\,dx = \frac{x}{x^2+y^2}\Big|_{x=0}^{x=1} = \frac{1}{1+y^2}; \qquad \int_0^1 \frac{dy}{1+y^2} = \arctan y\,\Big|_0^1 = \frac{\pi}{4};$$

$$\int_0^1 \frac{y^2 - x^2}{(x^2+y^2)^2}\,dy = \frac{y}{x^2+y^2}\Big|_{y=0}^{y=1} = -\frac{1}{x^2+1}; \qquad -\int_0^1 \frac{dx}{x^2+1} = -\arctan x\,\Big|_0^1 = -\frac{\pi}{4}.$$

8.2.5 Integration by Series Expansion, Special Non-Elementary Functions

It is not always possible to express an integral by elementary functions, even if the integrand is an elementary function. In many cases one can express these *non-elementary integrals* by series expansions. If the integrand can be expanded into a uniformly convergent series in the interval $[a, b]$, then one gets also a uniformly convergent series for the integral $\int_a^x f(t)\,dt$ by integrating it term by term.

1. Sine Integral ($|x| < \infty$, see also 14.4.3.2, **2.**, p. 756)

$$\mathrm{Si}\,(x) = \int_0^x \frac{\sin t}{t}\,dt = \frac{\pi}{2} - \int_x^\infty \frac{\sin t}{t}\,dt$$
$$= x - \frac{x^3}{3\cdot 3!} + \frac{x^5}{5\cdot 5!} - + \cdots + \frac{(-1)^n x^{2n+1}}{(2n+1)\cdot(2n+1)!} + \cdots. \tag{8.95}$$

2. Cosine Integral ($0 < x < \infty$)

$$\mathrm{Ci}\,(x) = -\int_x^\infty \frac{\cos t}{t}\,dt = C + \ln x - \int_0^x \frac{1-\cos t}{t}\,dt$$
$$= C + \ln x - \frac{x^2}{2\cdot 2!} + \frac{x^4}{4\cdot 4!} - + \cdots + \frac{(-1)^n x^{2n}}{2n\cdot(2n)!} + \cdots \quad \text{with} \tag{8.96a}$$
$$C = -\int_0^\infty e^{-t}\ln t\,dt = 0.577\,215\,665\ldots \qquad \text{(Euler constant)}. \tag{8.96b}$$

3. Integral Logarithm ($0 < x < 1$, for $1 < x < \infty$ as Cauchy Principal Value)

$$\mathrm{Li}\,(x) = \int_0^x \frac{dt}{\ln t} = C + \ln|\ln x| + \ln x + \frac{(\ln x)^2}{2\cdot 2!} + \cdots + \frac{(\ln x)^n}{n\cdot n!} + \cdots. \tag{8.97}$$

4. Exponential Integral ($-\infty < x < 0$, for $0 < x < \infty$ as Cauchy Principal Value)

$$\mathrm{Ei}\,(x) = \int_{-\infty}^{x} \frac{e^t}{t}\,dt = C + \ln|x| + x + \frac{x^2}{2\cdot 2!} + \cdots + \frac{x^n}{n\cdot n!} + \cdots. \tag{8.98a}$$

$$\mathrm{Ei}\,(\ln x) = \mathrm{Li}\,(x). \tag{8.98b}$$

5. Gauss Error Integral and Error Function

The *Gauss error integral* is defined for the domain $|x| < \infty$ and it is denoted by Φ. The following definitions and relations are valid:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}}\,dt, \tag{8.99a}$$

$$\lim_{x\to\infty} \Phi(x) = 1, \tag{8.99b}$$

$$\Phi_0(x) = \frac{1}{\sqrt{2\pi}} \int_{0}^{x} e^{-\frac{t^2}{2}}\,dt = \Phi(x) - \frac{1}{2}. \tag{8.99c}$$

The function $\Phi(x)$ is the distribution function of the standard normal distribution (see 16.2.4.2, p. 819) and its values are tabulated in **Table 21.17**, p. 1133.

The *error function* $\mathrm{erf}\,(x)$, often used in statistics (see also 16.2.4.2, p. 819), has a strong relation with the Gauss error integral:

$$\mathrm{erf}\,(x) = \frac{2}{\sqrt{\pi}} \int_{0}^{x} e^{-t^2}\,dt = 2\Phi_0(x\sqrt{2}), \tag{8.100a}$$

$$\lim_{x\to\infty} \mathrm{erf}\,(x) = 1, \tag{8.100b}$$

$$\mathrm{erf}\,(x) = \frac{2}{\sqrt{\pi}} \left(x - \frac{x^3}{1!\cdot 3} + \frac{x^5}{2!\cdot 5} - + \cdots + \frac{(-1)^n x^{2n+1}}{n!\cdot(2n+1)} + \cdots \right), \tag{8.100c}$$

$$\int_{0}^{x} \mathrm{erf}\,(t)\,dt = x\,\mathrm{erf}\,(x) + \frac{1}{\sqrt{\pi}}\left(e^{-x^2} - 1\right), \tag{8.100d}$$

$$\frac{d\,\mathrm{erf}\,(x)}{dx} = \frac{2}{\sqrt{\pi}} e^{-x^2}. \tag{8.100e}$$

6. Gamma Function and Factorial

1. Definition The *gamma function*, the Euler integral of the second kind (8.91), is an extension of the notion of factorial for arbitrary numbers x, even complex numbers, except zero and the negative integers. The curve of the function $\Gamma(x)$ is represented in **Fig. 8.23**. Its values are given in **Table 21.10**, p. 1105. It can be defined in two ways:

$$\Gamma(x) = \int_{0}^{\infty} e^{-t} t^{x-1}\,dt \quad (x > 0) \quad \text{or} \tag{8.101a}$$

$$\Gamma(x) = \lim_{n\to\infty} \frac{n^x \cdot n!}{x(x+1)(x+2)\dots(x+n)} \quad (x \neq 0, -1, -2, \dots). \tag{8.101b}$$

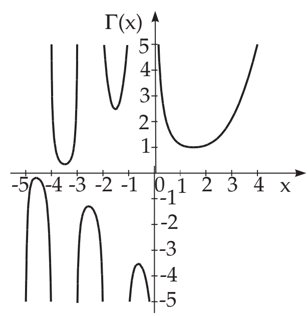


Figure 8.23

2. Properties of the Gamma Function

$$\Gamma(x+1) = x\Gamma(x), \tag{8.102a}$$

$$\Gamma(n+1) = n! \quad (n = 0, 1, 2, \dots), \tag{8.102b}$$

$$\Gamma(x)\,\Gamma(1-x) = \frac{\pi}{\sin \pi x} \quad (x \neq 0, \pm 1, \pm 2, \ldots), \tag{8.102c}$$

$$\Gamma\left(\frac{1}{2}\right) = 2\int_0^\infty e^{-t^2}\,dt = \sqrt{\pi}\,, \tag{8.102d}$$

$$\Gamma\left(n+\frac{1}{2}\right) = \frac{(2n)!\sqrt{\pi}}{n!2^{2n}} \qquad (n = 0, 1, 2, \ldots), \tag{8.102e}$$

$$\Gamma\left(-n+\frac{1}{2}\right) = \frac{(-1)^n n! 2^{2n}\sqrt{\pi}}{(2n)!} \quad (n = 0, 1, 2, \ldots). \tag{8.102f}$$

The formulas (8.102a) and (8.102c) are also valid for complex arguments z, but only if $\mathrm{Re}\,(z) > 0$ holds.

3. Generalization of the Notion of Factorial The notion of *factorial*, defined until now only for positive integers n (see 1.1.6.4, **3.**, p. 13), leads to the function

$$x! = \Gamma(x+1) \tag{8.103a}$$

as its extension for arbitrary real numbers. The following equalities are valid:

For positive integers x: $x! = 1 \cdot 2 \cdot 3 \cdots x$, (8.103b) for $x = 0$: $0! = \Gamma(1) = 1$, (8.103c)

for negative integers x: $x! = \pm\infty$, (8.103d) for $x = \frac{1}{2}$: $\left(\frac{1}{2}\right)! = \Gamma\left(\frac{3}{2}\right) = \frac{\sqrt{\pi}}{2}$, (8.103e)

for $x = -\frac{1}{2}$: $\left(-\frac{1}{2}\right)! = \Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$, (8.103f) for $x = -\frac{3}{2}$: $\left(-\frac{3}{2}\right)! = \Gamma\left(-\frac{1}{2}\right) = -2\sqrt{\pi}$. (8.103g)

An approximate determination of a factorial can be performed for numbers > 10, also for fractions n with the *Stirling formula*:

$$n! \approx \left(\frac{n}{e}\right)^n \sqrt{2\pi n}\left(1 + \frac{1}{12n} + \frac{1}{288n^2} + \cdots\right), \tag{8.103h}$$

$$\ln(n!) \approx \left(n + \frac{1}{2}\right)\ln n - n + \ln\sqrt{2\pi}. \tag{8.103i}$$

7. Elliptic Integrals

For the complete elliptic integrals (see 8.1.4.3, **2.**, p. 490) the following series expansions are valid:

$$K = \int_0^{\frac{\pi}{2}} \frac{d\vartheta}{\sqrt{1-k^2\sin^2\vartheta}} = \frac{\pi}{2}\left[1 + \left(\frac{1}{2}\right)^2 k^2 + \left(\frac{1\cdot 3}{2\cdot 4}\right)^2 k^4 + \left(\frac{1\cdot 3\cdot 5}{2\cdot 4\cdot 6}\right)^2 k^6 + \cdots\right],\ k^2 < 1, \tag{8.104}$$

$$E = \int_0^{\frac{\pi}{2}} \sqrt{1-k^2\sin^2\vartheta}\,d\vartheta = \frac{\pi}{2}\left[1 - \left(\frac{1}{2}\right)^2 \frac{k^2}{1} - \left(\frac{1\cdot 3}{2\cdot 4}\right)^2 \frac{k^4}{3} - \left(\frac{1\cdot 3\cdot 5}{2\cdot 4\cdot 6}\right)^2 \frac{k^6}{5} - \cdots\right],\quad k^2 < 1. \tag{8.105}$$

The numerical values of the elliptic integrals are given in **Table 21.9**, p. 1103.

8.3 Line Integrals

The notion of the integral can be generalized in different ways. While the domain of an ordinary definite integral is an interval on the numerical axis, for a *line integral*, the domain of integration is a segment of a planar or space curve. The curve, i.e., the path of integration can also be closed; it is called also

circuit integral and it gives the circulation of the function along the curve. There are distinguished line integrals of the first type, of the second type, or of general type.

8.3.1 Line Integrals of the First Type

8.3.1.1 Definitions

The line integral of the first type or *integral over an arc* is the definite integral

$$\int_{(C)} f(x,y)\,ds, \tag{8.106}$$

where $f(x,y)$ is a function of two variables defined on a connected domain and the integration is performed over an arc $C \equiv \overset{\frown}{AB}$ of a plane curve given by its equation. The considered arc is in the same domain, and it is called the *path of integration*. The numerical value of the line integral of the first type can be determined in the following way **(Fig. 8.24)**:

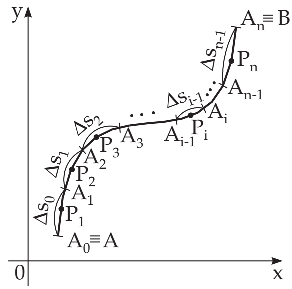


Figure 8.24

1. Decomposing the rectifiable arc segment $\overset{\frown}{AB}$ into n elementary parts by points $A_1, A_2, \dots, A_{n-1}$ chosen arbitrarily, starting at the initial point $A \equiv A_0$ and finishing at the endpoint $B \equiv A_n$.

2. Choosing arbitrary points P_i inside or at the end of the elementary arcs $\overset{\frown}{A_{i-1}A_i}$, with coordinates ξ_i and η_i.

3. Multiplying the values of the function $f(\xi_i, \eta_i)$ at the chosen points with the arc-length $\overset{\frown}{A_{i-1}A_i} = \Delta s_{i-1}$ which should be taken positive. (Since the arc is rectifiable, Δs_{i-1} is finite.)

4. Adding the n products $f(\xi_i, \eta_i)\Delta s_{i-1}$.

5. Evaluating the limit of the sum

$$\sum_{i=1}^{n} f(\xi_i, \eta_i)\Delta s_{i-1} \tag{8.107a}$$

as the arc-length of every elementary curve segment Δs_{i-1} tends to zero, while n obviously tends to ∞. If the limit of (8.107a) exists and is independent of the choice of the points A_i and P_i, then this limit is called the *line integral of the first type*, and the function $f(x,y)$ is called integrable along the curve C:

$$\int_{(C)} f(x,y)\,ds = \lim_{\substack{\Delta s_i \to 0 \\ n \to \infty}} \sum_{i=1}^{n} f(\xi_i, \eta_i)\Delta s_{i-1}. \tag{8.107b}$$

Analogously the line integral of the first type can be defined for a function $f(x,y,z)$ of three variables, whose path of integration is a curve segment of a space curve:

$$\int_{(C)} f(x,y,z)\,ds = \lim_{\substack{\Delta s_i \to 0 \\ n \to \infty}} \sum_{i=1}^{n} f(\xi_i, \eta_i, \zeta_i)\Delta s_{i-1}. \tag{8.107c}$$

8.3.1.2 Existence Theorem

The line integral of the first type (8.107b) or (8.107c) exists if the function $f(x,y)$ or $f(x,y,z)$ is continuous along the continuous arc segment C, and the curve has a tangent which varies continuously. In other words: The above limits exist and are independent of the choice of A_i and P_i. In this case, the functions $f(x,y)$ or $f(x,y,z)$ are called integrable along the curve C.

8.3.1.3 Evaluation of the Line Integral of the First Type

The calculation of the line integral of the first type can be done by reducing it to a definite integral.

1. The Equation of the Path of Integration is Given in Parametric Form

If the defining equations of the path are $x = x(t)$ and $y = y(t)$, then

$$\int_{(C)} f(x,y)\,ds = \int_{t_0}^{T} f[x(t),y(t)]\sqrt{[x'(t)]^2+[y'(t)]^2}\,dt \tag{8.108a}$$

holds, and in the case of a space curve $x = x(t)$, $y = y(t)$, and $z = z(t)$

$$\int_{(C)} f(x,y,z)\,ds = \int_{t_0}^{T} f[x(t),y(t),z(t)]\sqrt{[x'(t)]^2+[y'(t)]^2+[z'(t)]^2}\,dt, \tag{8.108b}$$

where t_0 is the value of the parameter t at the point A and T is the parameter value at B. The points A and B are chosen so that $t_0 < T$ holds.

2. The Equation of the Path of Integration is Given in Explicit Form y =y(x)

Substituting $t = x$ one gets from (8.108a) for the planar case

$$\int_{(C)} f(x,y)\,ds = \int_{a}^{b} f[x,y(x)]\sqrt{1+[y'(x)]^2}\,dx, \tag{8.109a}$$

and from (8.108b) for the three dimensional case

$$\int_{(C)} f(x,y,z)\,ds = \int_{a}^{b} f[x,y(x),z(x)]\sqrt{1+[y'(x)]^2+[z'(x)]^2}\,dx. \tag{8.109b}$$

Here a and b are the abscissae of the points A and B, where the relation $a < b$ must be fulfilled. Then one considers x as a parameter if every point corresponds to exactly one point on the projection of the curve segment C onto the x-axis, i.e., every point of the curve is uniquely determined by the value of its abscissa. If this condition does not hold, then the curve segment has to be partitioned into subsegments having this property. The line integral along the whole segment is equal to the sum of the line integrals along the subsegments.

8.3.1.4 Application of the Line Integral of the First Type

Some applications of the line integral of the first type are given in **Table 8.6**. The curve elements ds needed for the calculations of the line integrals are given for different coordinate systems in **Table 8.7**.

8.3.2 Line Integrals of the Second Type

8.3.2.1 Definitions

A *line integral of the second type* or an *integral over a projection* onto the x-, y- or z-axis is e.g., the definite integral

$$\int_{(C)} f(x,y)\,dx \tag{8.110a}$$

or

$$\int_{(C)} f(x,y,z)\,dx, \tag{8.110b}$$

where $f(x,y)$ or $f(x,y,z)$ are two or three variable functions defined on a connected domain, and the integration is done over a projection of a plane or space curve $C \equiv \overset{\frown}{AB}$ (given by its equation) onto the x-, y-, or z-axis. The path of integration is in the same domain.

The line integral of the second type one gets similarly to the line integral of the first type, but in the third step the values of the function $f(\xi_i,\eta_i)$ or $f(\xi_i,\eta_i,\zeta_i)$ are not multiplied by the arc-length of the elementary curve segments $\overset{\frown}{A_{i-1}A_i}$, but by its projections onto a coordinate axis **(Fig. 8.25)**.

Table 8.6 Line Integrals of the First Type

Length of a curve segment C	$L = \int\limits_{(C)} ds$
Mass of an inhomogeneous curve segment C	$M = \int\limits_{(C)} \varrho\, ds \quad (\varrho = f(x,y,z)$ density function)
Center of gravity coordinates	$x_C = \frac{1}{L}\int\limits_{(C)} x\varrho ds\,, \quad y_C = \frac{1}{L}\int\limits_{(C)} y\varrho ds\,, \quad z_C = \frac{1}{L}\int\limits_{(C)} z\varrho ds$
Moments of inertia of a plane curve in the x, y plane	$I_x = \int\limits_{(C)} x^2\varrho ds, \quad I_y = \int\limits_{(C)} y^2\varrho ds$
Moments of inertia of a space curve with respect to the coordinate axes	$I_x = \int\limits_{(C)} (y^2+z^2)\varrho ds, \quad I_y = \int\limits_{(C)} (x^2+z^2)\varrho ds,$ $I_z = \int\limits_{(C)} (x^2+y^2)\varrho ds$
In the case of homogeneous curves $\varrho = 1$ is substituted.	

Table 8.7 Curve Elements

Plane curve in the x, y plane	Cartesian coordinates $x, y = y(x)$	$ds = \sqrt{1 + [y'(x)]^2}dx$
	Polar coordinates $\varphi, \rho = \rho(\varphi)$ $x = \rho(\varphi)\cos\varphi, y = \rho(\varphi)\sin\varphi$	$ds = \sqrt{\rho^2(\varphi) + [\rho'(\varphi)]^2}d\varphi$
	Parametric form in Cartesian coordinates $x = x(t), y = y(t)$	$ds = \sqrt{[x'(t)]^2 + [y'(t)]^2}dt$
Space curve	Parametric form in Cartesian coordinates $x = x(t), y = y(t), z = z(t)$	$ds = \sqrt{[x'(t)]^2 + [y'(t)]^2 + [z'(t)]^2}dt$

1. Projection onto the x-Axis

With $\mathrm{Pr}_x\, \overset{\frown}{A_{i-1}A_i} = x_i - x_{i-1} = \Delta x_{i-1}$ one gets (8.111)

$$\int\limits_{(C)} f(x,y)\,dx = \lim_{\substack{\Delta x_{i-1}\to 0\\ n\to\infty}} \sum_{i=1}^{n} f(\xi_i,\eta_i)\,\Delta x_{i-1}, \tag{8.112a}$$

$$\int\limits_{(C)} f(x,y,z)\,dx = \lim_{\substack{\Delta x_{i-1}\to 0\\ n\to\infty}} \sum_{i=1}^{n} f(\xi_i,\eta_i,\zeta_i)\,\Delta x_{i-1}. \tag{8.112b}$$

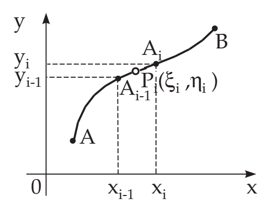


Figure 8.25

2. Projection onto the y-Axis

$$\int\limits_{(C)} f(x,y)\,dy = \lim_{\substack{\Delta y_{i-1}\to 0\\ n\to\infty}} \sum_{i=1}^{n} f(\xi_i,\eta_i)\,\Delta y_{i-1}, \tag{8.113a}$$

$$\int\limits_{(C)} f(x,y,z)\,dy = \lim_{\substack{\Delta y_{i-1}\to 0\\ n\to\infty}} \sum_{i=1}^{n} f(\xi_i,\eta_i,\zeta_i)\,\Delta y_{i-1}. \tag{8.113b}$$

3. Projection onto the z-Axis

$$\int\limits_{(C)} f(x,y,z)\,dz = \lim_{\substack{\Delta z_{i-1}\to 0\\ n\to\infty}} \sum_{i=1}^{n} f(\xi_i,\eta_i,\zeta_i)\,\Delta z_{i-1}. \tag{8.114}$$

8.3.2.2 Existence Theorem

The line integral of the second type in the form (8.112a), (8.113a), (8.112b), (8.113b) or (8.114) exists if the function $f(x,y)$ or $f(x,y,z)$ and also the curve are continuous along the arc segment C, and the curve has a continuously varying tangent there.

8.3.2.3 Calculation of the Line Integral of the Second Type

The calculation of the line integral of the second type can be done by reducing it to a definite integral.

1. The Path of Integration is Given in Parametric Form

With the parametric equations of the path of integration

$$x = x(t), \quad y = y(t) \quad \text{and (for a space curve)} \quad z = z(t) \tag{8.115}$$

we get the following formulas:

For (8.112a) $$\int\limits_{(C)} f(x,y)\,dx = \int\limits_{t_0}^{T} f[x(t),y(t)]x'(t)\,dt. \tag{8.116a}$$

For (8.113a) $$\int\limits_{(C)} f(x,y)\,dy = \int\limits_{t_0}^{T} f[x(t),y(t)]y'(t)\,dt. \tag{8.116b}$$

For (8.112b) $$\int\limits_{(C)} f(x,y,z)\,dx = \int\limits_{t_0}^{T} f[x(t),y(t),z(t)]x'(t)\,dt. \tag{8.116c}$$

For (8.113b) $$\int\limits_{(C)} f(x,y,z)\,dy = \int\limits_{t_0}^{T} f[x(t),y(t),z(t)]y'(t)\,dt. \tag{8.116d}$$

For (8.114) $$\int\limits_{(C)} f(x,y,z)\,dz = \int\limits_{t_0}^{T} f[x(t),y(t),z(t)]z'(t)\,dt. \tag{8.116e}$$

Here, t_0 and T are the values of the parameter t for the initial point A and the endpoint B of the arc segment. In contrast to the line integral of the first type, here we do not require the inequality $t_0 < T$.

Remark: In the case of reversion of the path of the integral, i.e., interchanging the points A and B, the sign of the integral changes.

2. The Path of Integration is Given in Explicit Form

In the case of a plane or space curve with the equations

$$y = y(x) \qquad \text{or} \qquad y = y(x),\; z = z(x) \tag{8.117}$$

as the path of integration, with the abscissae a and b of the points A and B, where the condition $a < b$ is no longer necessary, the abscissa x takes the place of the parameter t in the formulas (8.112a) – (8.114).

8.3.3 Line Integrals of General Type

8.3.3.1 Definition

A *line integral of general type* is the sum of the integrals of the second type along all the projections of a curve. If two functions $P(x,y)$ and $Q(x,y)$ of two variables, or three functions $P(x,y,z)$, $Q(x,y,z)$,

and $R(x, y, z)$ of three variables, are given along the given curve segment C, and the corresponding line integrals of the second type exist, then the following formulas are valid for a planar or for a space curve.

1. Planar Curve

$$\int_{(C)} (P\,dx + Q\,dy) = \int_{(C)} P\,dx + \int_{(C)} Q\,dy. \tag{8.118a}$$

2. Space Curve

$$\int_{(C)} (P\,dx + Q\,dy + R\,dz) = \int_{(C)} P\,dx + \int_{(C)} Q\,dy + \int_{(C)} R\,dz. \tag{8.118b}$$

The vector representation of the line integral of general type and an application of it in mechanics will be discussed in the chapter about vector analysis (see 13.3.1.1, p. 719).

8.3.3.2 Properties of the Line Integral of General Type

1. The Decomposition of the Path of the Integral

by a point M, which is on the curve, and it can even be outside of $\overset{\frown}{AB}$ **(Fig. 8.26)**, results in the decomposition of the integral into two parts:

$$\int_{\overset{\frown}{AB}} (P\,dx + Q\,dy) = \int_{\overset{\frown}{AM}} (P\,dx + Q\,dy) + \int_{\overset{\frown}{MB}} (P\,dx + Q\,dy).^{*} \tag{8.119}$$

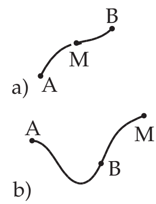


Figure 8.26

2. The Reverse of the Sense of the Path of Integration

changes the sign of the integral:

$$\int_{\overset{\frown}{AB}} (P\,dx + Q\,dy) = -\int_{\overset{\frown}{BA}} (P\,dx + Q\,dy).^{*} \tag{8.120}$$

3. Dependence on the Path

In general, the value of the line integral is dependent not only on the initial and endpoints but also on the path of integration **(Fig. 8.27)**:

$$\int_{\overset{\frown}{AMB}} (P\,dx + Q\,dy) \neq \int_{\overset{\frown}{ADB}} (P\,dx + Q\,dy).^{*} \tag{8.121}$$

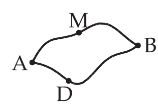


Figure 8.27

■ **A:** $I = \int_{(C)} (xy\,dx + yz\,dy + zx\,dz)$, where C is one turn of the helix $x = a\cos t,\ y = a\sin t, z = bt$ (see Helix on p. 260) from $t_0 = 0$ to $T = 2\pi$:

$$I = \int_0^{2\pi} (-a^3 \sin^2 t \cos t + a^2 bt \sin t \cos t + ab^2 t \cos t)\,dt = -\frac{\pi a^2 b}{2}.$$

■ **B:** $I = \int_{(C)} [y^2\,dx + (xy - x^2)\,dy]$, where C is the arc of the parabola $y^2 = 9x$ between the points

---

*Similar formulas are valid for the three-variable case.

$A(0,0)$ and $B(1,3)$: $I = \int_0^3 \left[\frac{2}{9}y^3 + \left(\frac{y^3}{9} - \frac{y^4}{81}\right)\right] dy = 6\frac{3}{20}$.

8.3.3.3 Integral Along a Closed Curve

1. Notion of the Integral Along a Closed Curve A *circuit integral* or the *circulation along a curve* is a line integral along a closed path of integration C, i.e., the initial point A and the end point B coincide. The following notation is used:

$$\oint_{(C)} (P\,dx + Q\,dy) \qquad \text{or} \qquad \oint_{(C)} (P\,dx + Q\,dy + R\,dz). \tag{8.122}$$

In general, this integral differs from zero. But it is equal to zero if the conditions (8.127) are satisfied, or if the integration is performed in a conservative field (see 13.3.1.6, p. 721). (See also zero-valued circulation, 13.3.1.6, p. 721.)

2. The Calculation of the Area *S* of a Plane Figure is a typical example of the application of the integral along a closed curve in the form

$$S = \frac{1}{2}\oint_{(C)} (x\,dy - y\,dx), \tag{8.123}$$

where C is the boundary curve of the plane figure. The integral is positive if the path is oriented counterclockwise.

8.3.4 Independence of the Line Integral of the Path of Integration

The condition for independence of a line integral of the path of integration is also called *integrability of the total differential.*

8.3.4.1 Two-Dimensional Case

If the line integral

$$\int_{(C)} [P(x,y)\,dx + Q(x,y)\,dy] \tag{8.124}$$

with continuous functions P and Q defined on a simple connected domain depends only on the initial point A and the endpoint B of the path of integration, and does not depend on the curve connecting these points, i.e., for arbitrary A and B and arbitrary paths of integration ACB and ADB **(Fig. 8.27)** the equality

$$\int_{\overset{\frown}{ACB}} (P\,dx + Q\,dy) = \int_{\overset{\frown}{ADB}} (P\,dx + Q\,dy) \tag{8.125}$$

holds, then it is a necessary and sufficient condition for the existence of a function $U(x,y)$ of two variables, whose total differential is the integrand of the line integral:

$$P\,dx + Q\,dy = d\,U, \quad (8.126\text{a}) \qquad \text{i.e.,} \quad P = \frac{\partial U}{\partial x}, \quad Q = \frac{\partial U}{\partial y}. \quad (8.126\text{b})$$

The function $U(x,y)$ is a primitive function of the total differential (8.126a). In physics, the primitive function $U(x,y)$ means the potential in a vector field (see 13.3.1.6, **4.**, p. 722).

8.3.4.2 Existence of a Primitive Function

A necessary and sufficient criterion for the existence of the *primitive function*, the *integrability condition* for the expression $P\,dx + Q\,dy$, is the equality of the partial derivatives

$$\frac{\partial P}{\partial y} = \frac{\partial Q}{\partial x}, \tag{8.127}$$

where also the continuity of the partial derivatives is required.

8.3.4.3 Three-Dimensional Case

The condition of independence of the line integral

$$\int [P(x,y,z)\,dx + Q(x,y,z)\,dy + R(x,y,z)\,dz] \tag{8.128}$$

of the path of integration analogously to the two-dimensional case is the existence of a primitive function $U(x,y,z)$ for which

$$P\,dx + Q\,dy + R\,dz = d\,U, \tag{8.129a}$$

holds, i.e.,

$$P = \frac{\partial U}{\partial x}, \quad Q = \frac{\partial U}{\partial y}, \quad R = \frac{\partial U}{\partial z}. \tag{8.129b}$$

The integrability condition is now that the three equalities for the partial derivatives

$$\frac{\partial Q}{\partial z} = \frac{\partial R}{\partial y}, \quad \frac{\partial R}{\partial x} = \frac{\partial P}{\partial z}, \quad \frac{\partial P}{\partial y} = \frac{\partial Q}{\partial x} \tag{8.129c}$$

should be simultaneously satisfied, provided that the partial derivatives are continuous.

■ The work W (see also 8.2.2.3, **2.**, p. 504) is defined as the scalar product of force $\vec{F}(\vec{r})$ and displacement $\vec{s}$. In a conservative field the work depends only on the place $\vec{r}$, but not on the velocity $\vec{v}$. With $\vec{F} = P\vec{e}_x + Q\vec{e}_y + R\vec{e}_z = \text{grad}V$ and $\vec{ds} = dx\vec{e}_x + dy\vec{e}_y + dz\vec{e}_z$ the relations (8.129a), (8.129b) are satisfied for the potential $V(\vec{r})$, and (8.129c) is valid. Independently of the path between the points P_1 and P_2 holds:

$$W = \int_{P_1}^{P_2} \vec{F}(\vec{r}) \cdot \vec{ds} = \int_{P_1}^{P_2} [P\,dx + Q\,dy + R\,dz] = V(P_2) - V(P_1). \tag{8.130}$$

8.3.4.4 Determination of the Primitive Function

1. Two-Dimensional Case (Fig.8.28)

If the integrability condition (8.127) is satisfied, then along an arbitrary path of integration connecting an arbitrary fixed point $A(x_0, y_0)$ with the variable point $P(x,y)$ and passing through the domain where (8.127) is valid, the primitive function $U(x,y)$ is equal to the line integral

$$U = \int\limits_{\widehat{AP}} (P\,dx + Q\,dy). \tag{8.131}$$

In practice, it is convenient to choose a path of integration parallel to the coordinate axes, i.e., one of the segments AKP or ALP, if they are inside the domain where (8.127) is valid. There exist two formulas for the calculation of the primitive function $U(x,y)$ and the total differential $P\,dx + Q\,dy$:

$$U = U(x_0, y_0) + \int\limits_{\overline{AK}} + \int\limits_{\overline{KP}} = C + \int\limits_{x_0}^{x} P(\xi, y_0)\,d\xi + \int\limits_{y_0}^{y} Q(x, \eta)\,d\eta\,, \tag{8.132a}$$

$$U = U(x_0, y_0) + \int\limits_{\overline{AL}} + \int\limits_{\overline{LP}} = C + \int\limits_{y_0}^{y} Q(x_0, \eta)\,d\eta + \int\limits_{x_0}^{x} P(\xi, y)\,d\xi\,. \tag{8.132b}$$

Here C is an arbitrary constant.

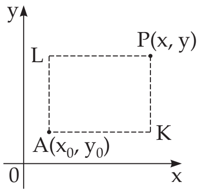


Figure 8.28

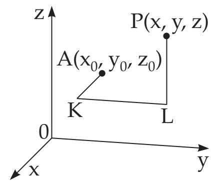


Figure 8.29

2. Three-Dimensional Case (Fig. 8.29)

If the condition (8.129c) is satisfied, the primitive function can be calculated for the path of integration $AKLP$ with the formulas

$$U = U(x_0, y_0, z_0) + \int\limits_{AK} + \int\limits_{KL} + \int\limits_{LP}$$

$$= \int\limits_{x_0}^{x} P(\xi, y_0, z_0)\, d\xi + \int\limits_{y_0}^{y} Q(x, \eta, z_0)\, d\eta + \int\limits_{z_0}^{z} R(x, y, \xi)\, d\xi + C \quad (C \text{ arbitrary constant}). \tag{8.133}$$

For the other five possibilities of a path of integration with the segments being parallel to the coordinate axes one gets five further formulas.

■ **A:** $P\,dx + Q\,dy = -\dfrac{y\,dx}{x^2+y^2} + \dfrac{x\,dy}{x^2+y^2}$. The condition (8.129c) is satisfied: $\dfrac{\partial P}{\partial y} = \dfrac{\partial Q}{\partial x} = \dfrac{y^2-x^2}{(x^2+y^2)^2}$. Application of the formula (8.132b) and the substitution of $x_0 = 0,\ y_0 = 1$ ($x_0 = 0,\ y_0 = 0$ may not be chosen since the functions P and Q are discontinuous at the point $(0,0)$) resulting in $U = \displaystyle\int_1^y \frac{0\cdot d\eta}{0^2+\eta^2} + \int_0^x \frac{-y\,d\xi}{\xi^2+y^2} + U(0,1) = -\arctan\frac{x}{y} + C = \arctan\frac{y}{x} + C_1.$

■ **B:** $P\,dx + Q\,dy + R\,dz = z\left(\dfrac{1}{x^2y} - \dfrac{1}{x^2+z^2}\right)dx + \dfrac{z}{xy^2}\,dy + \left(\dfrac{x}{x^2+z^2} - \dfrac{1}{xy}\right)dz$. The relations (8.129c) are satisfied. Application of the formula (8.133) and substitution of $x_0 = 1,\ y_0 = 1,\ z_0 = 0$ result in $U = \displaystyle\int_1^x 0\cdot d\xi + \int_1^y 0\cdot d\eta + \int_0^z \left(\frac{x}{x^2+\zeta^2} - \frac{1}{xy}\right) d\zeta + C = \arctan\frac{z}{x} - \frac{z}{xy} + C.$

8.3.4.5 Zero-Valued Integral Along a Closed Curve

The integral along a closed curve, i.e., the line integral $P\,dx + Q\,dy$ is equal to zero, if the relation (8.127) is satisfied, and if there is no point inside the curve where even one of the functions $P,\ Q,\ \dfrac{\partial P}{\partial y}$ or $\dfrac{\partial Q}{\partial x}$ is discontinuous or not defined.

Remark: The value of the integral can be equal to zero also without this conditions, but then one gets this value only after performing the corresponding calculations.

8.4 Multiple Integrals

The notion of the integral can be extended to higher dimensions. If the domain of integration is a region in the plane or on a surface in space, then the integral is called a *surface integral*, if the domain is a part of space, then it is called a *volume integral*. Furthermore, for the different special applications there are other special notations.

8.4.1 Double Integrals

8.4.1.1 Notion of the Double Integral

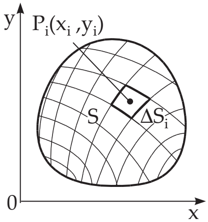


Figure 8.30

1. Definition

The *double integral* of a function of two variables $u = f(x,y)$ over a planar domain S in the x, y plane is denoted by

$$\int_S f(x,y)\,dS = \iint_S f(x,y)\,dy\,dx. \tag{8.134}$$

It is a number, if it exists, and it is defined in the following way **(Fig. 8.30)**:

1. Decomposition of the domain S into n elementary domains ΔS_i.

2. Choosing an arbitrary point $P_i(x_i, y_i)$ in the interior or on the boundary of every elementary domain.

3. Multiplication of the value of the function $u = f(x_i, y_i)$ at this point by the area ΔS_i of the corresponding elementary domain.

4. Summation of these products $f(x_i, y_i)\Delta S_i$.

5. Calculation of the limit of the sum

$$\sum_{i=1}^{n} f(x_i, y_i)\Delta S_i \tag{8.135a}$$

as the *diameter of the elementary domains* tends to zero, consequently ΔS_i tends to zero, and so n tends to ∞. (The diameter of a set of points is the supremum of the distances between the points of the set.) The requirement ΔS tends to zero is not enough, because, e.g., in the case of a rectangle the area can be close to zero also if only one side is small and the other is not, so the considered points could be far from each other. If this limit exists independently of the partition of the domain S into elementary domains and also of the choice of the points $P_i(x_i, y_i)$, then it is called the double integral of the function $u = f(x,y)$ over the domain S, the domain of integration, and one writes:

$$\int_S f(x,y)\,dS = \lim_{\substack{\Delta S_i \to 0 \\ n\to\infty}} \sum_{i=1}^{n} f(x_i, y_i)\,\Delta S_i. \tag{8.135b}$$

2. Existence Theorem

If the function $f(x,y)$ is continuous on the domain of integration including the boundary, then the double integral (8.135b) exists. (This condition is sufficient but not necessary.)

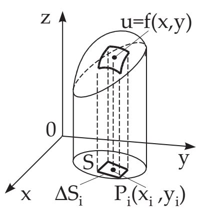


Figure 8.31

3. Geometrical Meaning

The geometrical meaning of the double integral is the volume of a solid whose base is the domain in the x, y plane, whose side is a cylindrical surface with generators parallel to the z-axis, and it is bounded above by the surface defined by $u = f(x,y)$ **(Fig. 8.31)**. Every term $f(x_i, y_i)\Delta S_i$ of the sum (8.135b) corresponds to an elementary cell of a prism with base ΔS_i and with altitude $f(x_i, y_i)$. The sign of the volume is positive or negative, according to whether the considered part of the surface $u = f(x,y)$ is above or under the x, y plane. If the surface intersects the x, y plane, then the volume is the algebraic sum of the positive and negative parts.

If the value of the function is identically 1 ($f(x,y) \equiv 1$), then the volume has the numerical value of the area of the domain S in the x, y plane.

8.4.1.2 Evaluation of the Double Integral

The evaluation of the double integral is reduced to the evaluation of a repeated integral, i.e., to the evaluation of two consecutive integrals.

1. Evaluation in Cartesian Coordinates

If the double integral exists, then one can consider any type of partition of the domain of integration, such as a partition into rectangles. After dividing the domain of integration into infinitesimal rectangles by coordinate lines **(Fig. 8.32a)** it follows the calculation of the sum of all differentials $f(x, y)dS$ starting with all the rectangles along every vertical stripe, then along every horizontal stripe. (The interior sum is an integral approximation sum with respect to the variable y, the exterior one with respect to x.) If the integrand is continuous, then this repeated integral is equal to the double integral on this domain. The analytic notation is:

$$\int_S f(x,y)\,dS = \int_a^b \left[\int_{\varphi_1(x)}^{\varphi_2(x)} f(x,y)\,dy \right] dx = \int_a^b \int_{\varphi_1(x)}^{\varphi_2(x)} f(x,y)\,dy\,dx. \tag{8.136a}$$

Here $y = \varphi_2(x)$ and $y = \varphi_1(x)$ are the equations of the upper and lower boundary curves $(\overset{\frown}{AB})_{\text{above}}$ and $(\overset{\frown}{AB})_{\text{below}}$ of the surface region S ($\varphi_1 \le \varphi_2$; φ_1, φ_2 continuous). Here a and b are the abscissae of the points of the curves to the very left and to the very right. The elementary area in Cartesian coordinates is

$$dS = dx\,dy. \tag{8.136b}$$

(The area of the rectangle is $\Delta x \Delta y$ independently of the value of x.) For the first integration x is

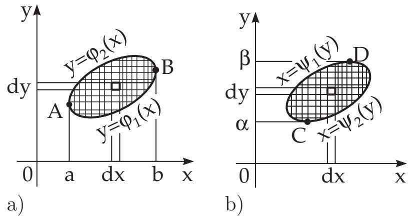


Figure 8.32

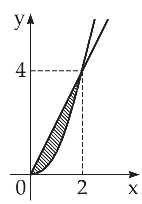


Figure 8.33

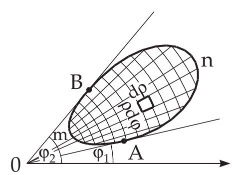


Figure 8.34

handled as a constant. The square brackets in (8.136a) can be omitted, since according to the notation the interior integral is referred to the interior integration variable, the exterior integral is referred to the second variable. In (8.136a) the differential signs dx and dy are at the end of the integrand. It is also usual to put these signs right after the corresponding integral signs, in front of the integrand.

The summation can be performed in reversed order, too, **(Fig. 8.32b)**. If the integrand is continuous, then it results also in the double integral:

$$\int_S f(x,y)\,dS = \int_\alpha^\beta \int_{\psi_1(y)}^{\psi_2(y)} f(x,y)\,dx\,dy. \tag{8.136c}$$

■ Calculation of $A = \int_S xy^2\,dS$, where S is the surface region between the parabola $y = x^2$ and the line $y = 2x$ **(Fig. 8.33)** as $A = \int_0^2 \int_{x^2}^{2x} xy^2\,dy\,dx = \int_0^2 x\,dx \left[\frac{y^3}{3}\right]_{x^2}^{2x} = \frac{1}{3}\int_0^2 (8x^4 - x^7)\,dx = \frac{32}{5}$ or

$$A = \int_0^4 \int_{y/2}^{\sqrt{y}} xy^2\,dx\,dy = \int_0^2 y^2\,dy \left[\frac{x^2}{2}\right]_{y/2}^{\sqrt{y}} = \frac{1}{2}\int_0^4 y^2 \left(y - \frac{y^2}{4}\right) dy = \frac{32}{5}.$$

2. Evaluation in Polar Coordinates

The integration domain is divided by coordinate lines into elementary parts bounded by the arcs of two concentric circles and two segments of rays issuing from the pole **(Fig. 8.34)**. The area of the

elementary domain in polar coordinates has the form

$$dS = \rho\, d\rho\, d\varphi. \tag{8.137a}$$

(The area of an elementary part determined by the same $\Delta\rho$ and $\Delta\varphi$ is obviously smaller being close to the origin, and larger far from it.) With an integrand given in polar coordinates $w = f(\rho,\varphi)$ a summation is to be performed first along each sector, then with respect to all sectors:

$$\int_S f(\rho,\varphi)\, dS = \int_{\varphi_1}^{\varphi_2} \int_{\rho_1(\varphi)}^{\rho_2(\varphi)} f(\rho,\varphi)\,\rho\, d\rho\, d\varphi\,. \tag{8.137b}$$

Here $\rho = \rho_1(\varphi)$ and $\rho = \rho_2(\varphi)$ are the equations of the interior and the exterior boundary curves $(\overset{\frown}{AmB})$ and $(\overset{\frown}{AnB})$ of the surface S and φ_1 and φ_2 are the infimum and supremum of the polar angles of the points of the domain. The reverse order of integration is seldom used.

■ Calculation of the special integral $A = \int_S \rho \sin^2\varphi\, dS$, where S is the surface of the half-circle with $\rho = 3\cos\varphi$ $(0 \le \varphi \le \pi/2)$ (see **Fig. 8.35**):

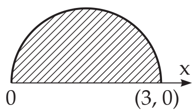


Figure 8.35

$$A = \int_0^{\pi/2}\int_0^{3\cos\varphi} \rho^2 \sin^2\varphi\, d\rho\, d\varphi = \int_0^{\pi/2} \sin^2\varphi\, d\varphi \left[\frac{\rho^3}{3}\right]_0^{3\cos\varphi}$$
$$= 9\int_0^{\pi/2} \sin^2\varphi \cos^3\varphi\, d\varphi = \frac{6}{5}\,.$$

3. Evaluation with Arbitrary Curvilinear Coordinates u and v

The coordinates are defined by the relations

$$x = x(u,v), \quad y = y(u,v) \tag{8.138}$$

(see 3.6.3.1, p. 261). The domain of integration is partitioned by coordinate lines $u = $ const and $v = $ const into infinitesimal surface elements **(Fig. 8.36)** and the integrand is expressed by the coordinates u and v. Performing the summation along one strip, e.g., along $v = $ const, then over all strips, yields

$$\int_S f(u,v)\, dS = \int_{u_1}^{u_2}\int_{v_1(u)}^{v_2(u)} f(u,v)|D|\, dv\, du. \tag{8.139}$$

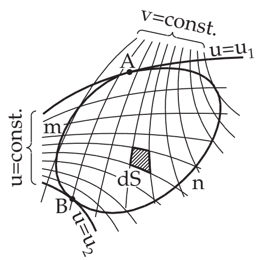


Figure 8.36

Here $v = v_1(u)$ and $v = v_2(u)$ are the equations of the boundary curves $\overset{\frown}{AmB}$ and $\overset{\frown}{AnB}$ of the surface S. u_1 and u_2 denote the infimum and the supremum of the values of u of the points belonging to the surface S. $|D|$ denotes the absolute value of the *Jacobian determinant* (*functional determinant*)

$$D = \frac{D(x,y)}{D(u,v)} = \begin{vmatrix} \dfrac{\partial x}{\partial u} & \dfrac{\partial x}{\partial v} \\ \dfrac{\partial y}{\partial u} & \dfrac{\partial y}{\partial v} \end{vmatrix}. \tag{8.140a}$$

The area of the elementary domain dS in curvilinear coordinates can be easily expressed:

$$dS = |D|\, dv\, du\,. \tag{8.140b}$$

The formula (8.137b) is a special case of (8.139) for the polar coordinates $x = \rho\cos\varphi,\ y = \rho\sin\varphi$. The functional determinant here is $D = \rho$.

The curvilinear coordinates are chosen so that the limits of integration in the formula (8.139) are as simple as possible, and also the integrand is not very complicated.

■ Calculation of $A = \int_S f(x,y)\,dS$ for the case when S is the interior of an asteroid (see 2.13.4, p. 104), with $x = a\cos^3 t,\ y = a\sin^3 t$ **(Fig. 8.37)**. First the curvilinear coordinates u and v as $x = u\cos^3 v,\ y = u\sin^3 v$ are introduced whose coordinate lines $u = c_1$ represents a family of similar asteroids with equations $x = c_1\cos^3 v$ and $y = c_1\sin^3 v$. The coordinate lines $v = c_2$ are rays with the equations $y = kx$, where $k = \tan^3 c_2$ holds. This gives

$$D = \begin{vmatrix} \cos^3 v & -3u\cos^2 v\sin v \\ \sin^3 v & 3u\sin^2 v\cos v \end{vmatrix} = 3u\sin^2 v\cos^2 v,$$

$$A = \int_0^a \int_0^{2\pi} f(x(u,v), y(u,v))\, 3u\sin^2 v\cos^2 v\, dv\, du.$$

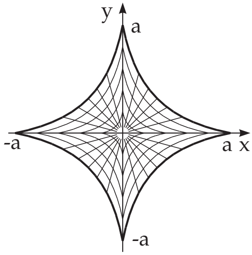


Figure 8.37

Tabelle 8.8 Plane Elements of Area

Coordinates	**Element of Area**
Cartesian coordinates x, y	$dS = dy\,dx$
Polar coordinates ρ, φ	$dS = \rho\,d\rho\,d\varphi$
Arbitrary curvilinear coordinates u, v	$dS = \|D\|\,du\,dv$ (D Jacobian determinant)

8.4.1.3 Applications of the Double Integral

Some applications of the double integral are collected in **Table 8.9**, p. 528. The required areas of elementary domains in Cartesian and polar coordinates are given in **Table 8.8**.

8.4.2 Triple Integrals

The *triple integral* is an extension of the notion of the integral into three-dimensional domains. It also is called *volume integral*.

8.4.2.1 Notion of the Triple Integral

1. Definition

One defines the triple integral of a function $f(x,y,z)$ of three variables over a three-dimensional domain V analogously to the definition of the double integral. One writes:

$$\int_V f(x,y,z)\,dV = \iiint_V f(x,y,z)\,dz\,dy\,dx. \tag{8.141}$$

The volume V **(Fig. 8.38)** is partitioned into elementary volumes ΔV_i. Then the products $f(x_i,y_i,z_i)$ ΔV_i are formed, where the point $P_i(x_i,y_i,z_i)$ is inside the elementary volume or it is on the boundary. The triple integral is the limit of the sum of these products with all the elementary volumes in which the volume V is partitioned, then the diameter of every elementary volume tends to zero, i.e., their number tends to ∞. The triple integral exists only if the limit is independent of the partition into elementary volumes and the choice of the points $P_i(x_i,y_i,z_i)$. Then holds:

$$\int_V f(x,y,z)\,dV = \lim_{\substack{\Delta V_i \to 0 \\ n\to\infty}} \sum_{i=1}^{n} f(x_i,y_i,z_i)\,\Delta V_i. \tag{8.142}$$

Table 8.9 Applications of the Double Integral

General Formula	Cartesian Coordinates	Polar Coordinates
1. Area of a plane figure:		
$S = \int\limits_S dS$	$= \iint dy\,dx$	$= \iint \rho\,d\rho\,d\varphi$
2. Surface:		
$S_O = \int\limits_S \frac{dS}{\cos\gamma}$	$= \iint \sqrt{1+\left(\frac{\partial z}{\partial x}\right)^2+\left(\frac{\partial z}{\partial y}\right)^2}\,dy\,dx$	$= \iint \sqrt{\rho^2+\rho^2\left(\frac{\partial z}{\partial \rho}\right)^2+\left(\frac{\partial z}{\partial \varphi}\right)^2}\,d\rho\,d\varphi$
3. Volume of a cylinder:		
$V = \int\limits_S z\,dS$	$= \iint z\,dy\,dx$	$= \iint z\rho\,d\rho\,d\varphi$
4. Moment of inertia of a plane figure, with respect to the *x*-axis:		
$I_x = \int\limits_S y^2\,dS$	$= \iint y^2\,dy\,dx$	$= \iint \rho^3\sin^2\varphi\,d\rho\,d\varphi$
5. Moment of inertia of a plane figure, with respect to the pole 0:		
$I_0 = \int\limits_S \rho^2\,dS$	$= \iint (x^2+y^2)\,dy\,dx$	$= \iint \rho^3\,d\rho\,d\varphi$
6. Mass of a plane figure with the density function ϱ:		
$M = \int\limits_S \varrho\,dS$	$= \iint \varrho\,dy\,dx$	$= \iint \varrho\rho\,d\rho\,d\varphi$
7. Coordinates of the center of gravity of a homogeneous plane figure:		
$x_C = \frac{\int\limits_S x\,dS}{S}$	$= \frac{\iint x\,dy\,dx}{\iint dy\,dx}$	$= \frac{\iint \rho^2\cos\varphi\,d\rho\,d\varphi}{\iint \rho\,d\rho\,d\varphi}$
$y_C = \frac{\int\limits_S y\,dS}{S}$	$= \frac{\iint y\,dy\,dx}{\iint dy\,dx}$	$= \frac{\iint \rho^2\sin\varphi\,d\rho\,d\varphi}{\iint \rho\,d\rho\,d\varphi}$

2. Existence Theorem

The existence theorem for the triple integral is a perfect analogue of the existence theorem for the double integral.

8.4.2.2 Evaluation of the Triple Integral

The evaluation of triple integrals is reduced to repeated evaluation of three ordinary integrals. If the triple integral exists, then one can consider any partition of the domain of integration.

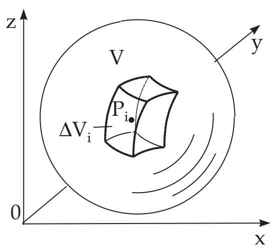


Figure 8.38

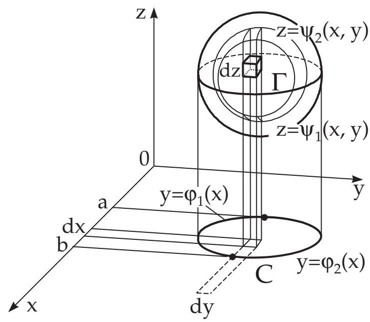


Figure 8.39

1. Evaluation in Cartesian Coordinates

The domain of integration can be considered as a volume V here. A decomposition of the domain is formed by coordinate surfaces, in this case by planes, into infinitesimal parallelepipeds, i.e., their diameter is an infinitesimal quantity **(Fig. 8.39)**. Then one performs the summation of all the products $f(x, y, z)\,dV$, starting the summation along the vertical columns, i.e., summation with respect to z, then in all columns of one slice, i.e., summation with respect to y, and finally in all such slices, i.e., summation with respect to x. Every single sum for any column is an approximation sum of an integral, and if the diameter of the parallelepipeds tends to zero, then the sums tend to the corresponding integrals, and if the integrand is continuous, then this repeated integral is equal to the triple integral. Analytically:

$$\int\limits_V f(x,y,z)\,dV = \int\limits_a^b \left\{ \int\limits_{\varphi_1(x)}^{\varphi_2(x)} \left[\int\limits_{\psi_1(x,y)}^{\psi_2(x,y)} f(x,y,z)\,dz \right] dy \right\} dx$$

$$= \int\limits_a^b \int\limits_{\varphi_1(x)}^{\varphi_2(x)} \int\limits_{\psi_1(x,y)}^{\psi_2(x,y)} f(x,y,z)\,dz\,dy\,dx. \tag{8.143a}$$

Here $z = \psi_1(x, y)$ and $z = \psi_2(x, y)$ are the equations of the lower and upper part of the surface bounding the domain of integration V (see limiting curve Γ in **Fig. 8.39**); $dx\,dy\,dz$ is the elementary volume in the Cartesian coordinate system. $y = \varphi_1(x)$ and $y = \varphi_2(x)$ are the functions describing the lower and upper part of the curve C which is the boundary line of the projection of the volume onto the x, y plane, and $x = a$ and $x = b$ are the extreme values of the x coordinates of the points of the volume under consideration (and also the projection under consideration). There are the following postulates for the domain of integration: The functions $\varphi_1(x)$ and $\varphi_2(x)$ are defined and continuous in the interval $a \le x \le b$, and they satisfy the inequality $\varphi_1(x) \le \varphi_2(x)$. The functions $\psi_1(x, y)$ and $\psi_2(x, y)$ are defined and continuous on the domain $a \le x \le b$, $\varphi_1(x) \le y \le \varphi_2(x)$, and also $\psi_1(x, y) \le \psi_2(x, y)$ holds. In this way, every point (x, y, z) in V satisfies the relations

$$a \le x \le b, \qquad \varphi_1(x) \le y \le \varphi_2(x), \qquad \psi_1(x,y) \le z \le \psi_2(x,y). \tag{8.143b}$$

Just as with double integrals, the order of integration can be changed, then the limiting functions will change in the same sense. (Formally: the limits of the outermost integral must be constants, and any limit may contain variables only of exterior integrals.)

■ Calculate the integral $I = \int\limits_V (y^2 + z^2)\,dV$ for a pyramid bounded by the coordinate planes and the plane $x + y + z = 1$:

$I = \int_0^1 \int_0^{1-x} \int_0^{1-x-y} (y^2 + z^2)\,dz\,dy\,dx = \int_0^1 \left\{ \int_0^{1-x} \left[\int_0^{1-x-y} (y^2 + z^2)\,dz \right] dy \right\} dx = \frac{1}{30}\,.$

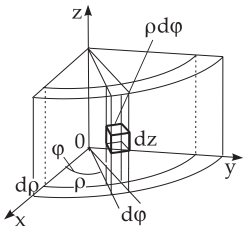


Figure 8.40

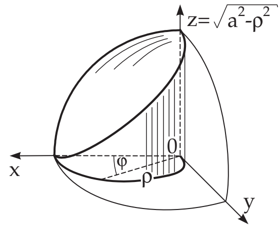


Figure 8.41

2. Evaluation in Cylindrical Coordinates

The domain of integration is decomposed into infinitesimal elementary cells by coordinate surfaces $\rho = \text{const}$, $\varphi = \text{const}$, $z = \text{const}$ **(Fig. 8.40)**. The *volume of an elementary domain in cylindrical coordinates* (**Table 8.10**, p. 532) is

$$dV = \rho\,dz\,d\rho\,d\varphi. \tag{8.144a}$$

After defining the integrand by cylindrical coordinates $f(\rho, \varphi, z)$ the integral is:

$$\int\limits_V f(\rho, \varphi, z)\,dV = \int\limits_{\varphi_1}^{\varphi_2} \int\limits_{\rho_1(\varphi)}^{\rho_2(\varphi)} \int\limits_{z_1(\rho,\varphi)}^{z_2(\rho,\varphi)} f(\rho, \varphi, z)\,\rho\,dz\,d\rho\,d\varphi. \tag{8.144b}$$

■ Calculate the integral $I = \int\limits_V dV$ for a solid **(Fig. 8.41)** bounded by the x, y plane, the x, z plane, the cylindrical surface $x^2 + y^2 = ax$ and the sphere $x^2 + y^2 + z^2 = a^2$: $z_1 = 0$, $z_2 = \sqrt{a^2 - x^2 - y^2} = \sqrt{a^2 - \rho^2}$; $\rho_1 = 0$, $\rho_2 = a\cos\varphi$; $\varphi_1 = 0$, $\varphi_2 = \frac{\pi}{2}$. $I = \int_0^{\pi/2} \int_0^{a\cos\varphi} \int_0^{\sqrt{a^2-\rho^2}} \rho\,dz\,d\rho\,d\varphi = \int_0^{\pi/2} \left\{ \int_0^{a\cos\varphi} \left[\int_0^{\sqrt{a^2-\rho^2}} dz \right] \rho\,d\rho \right\} d\varphi = \frac{a^3}{18}(3\pi - 4)$. Since $f(\rho, \varphi, z) = 1$, the integral is equal to the volume of the solid.

3. Evaluation in Spherical Coordinates

The domain of integration is decomposed into infinitesimal elementary cells by coordinate surfaces $r = \text{const}$, $\varphi = \text{const}$, $\vartheta = \text{const}$ **(Fig. 8.42)**. The *volume of an elementary domain in spherical coordinates* (see **Table 8.10**, p. 532) is

$$dV = r^2 \sin\vartheta\,dr\,d\vartheta\,d\varphi. \tag{8.145a}$$

For the integrand $f(r, \varphi, \vartheta)$ in spherical coordinates, the integral is:

$$\int\limits_V f(r, \varphi, \vartheta)\,dV = \int\limits_{\varphi_1}^{\varphi_2} \int\limits_{\vartheta_1(\varphi)}^{\vartheta_2(\varphi)} \int\limits_{r_1(\vartheta,\varphi)}^{r_2(\vartheta,\varphi)} f(r, \varphi, \vartheta)\,r^2 \sin\vartheta\,dr\,d\vartheta\,d\varphi. \tag{8.145b}$$

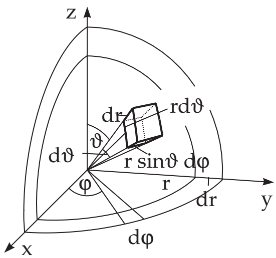


Figure 8.42

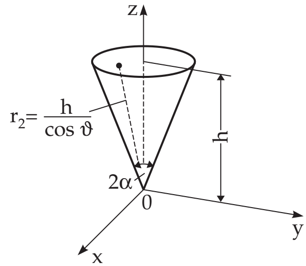


Figure 8.43

■ Calculate the integral $I = \int_V \frac{\cos\vartheta}{r^2}\,dV$ for a cone whose vertex is at the origin, and its symmetry axis is the z-axis. The angle at the vertex is 2α, the altitude of the cone is h **(Fig. 8.43)**. Consequently holds: $r_1 = 0, r_2 = \dfrac{h}{\cos\vartheta}$; $\quad \vartheta_1 = 0, \vartheta_2 = \alpha$; $\quad \varphi_1 = 0, \varphi_2 = 2\pi$.

$$I = \int_0^{2\pi}\int_0^{\alpha}\int_0^{h/\cos\vartheta} \frac{\cos\vartheta}{r^2} r^2 \sin\vartheta\,dr\,d\vartheta\,d\varphi = \int_0^{2\pi}\left\{\int_0^{\alpha}\cos\vartheta\sin\vartheta\left[\int_0^{h/\cos\vartheta} dr\right]d\vartheta\right\}d\varphi$$
$$= 2\pi\,h\,(1-\cos\alpha).$$

4. Evaluation in Arbitrary Curvilinear Coordinates u, v, w

The coordinates are defined by the equations

$$x = x(u,v,w), \qquad y = y(u,v,w), \qquad z = z(u,v,w) \tag{8.146}$$

(see 3.6.3.1, p. 261). The domain of integration is decomposed into infinitesimal elementary cells by the coordinate surfaces $u = \text{const}$, $v = \text{const}$, $w = \text{const}$. The *volume of an elementary domain in arbitrary coordinates* (see **Table 8.10**, p. 532) is:

$$dV = |D|\,du\,dv\,dw, \qquad \text{with} \qquad D = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} & \frac{\partial x}{\partial w} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} & \frac{\partial y}{\partial w} \\ \frac{\partial z}{\partial u} & \frac{\partial z}{\partial v} & \frac{\partial z}{\partial w} \end{vmatrix}, \tag{8.147a}$$

i.e., D is the Jacobian determinant. For the integrand $f(u,v,w)$ in curvilinear coordinates u, v, w, the integral is:

$$\int_V f(u,v,w)\,dV = \int_{u_1}^{u_2}\int_{v_1(u)}^{v_2(u)}\int_{w_1(u,v)}^{w_2(u,v)} f(u,v,w)\,|D|\,dw\,dv\,du. \tag{8.147b}$$

Remark: The formulas (8.144b) and (8.145b) are special cases of (8.147b).
For cylindrical coordinates $D = \rho$ holds, for spherical coordinates $D = r^2\sin\vartheta$ is valid.
If the integrand is continuous, then one can change the order of integration in any coordinate system.
A curvilinear coordinate system is chosen such that the determination of the limits of the integral (8.147b), and also the calculation of the integral, should be as easy as possible.

8.4.2.3 Applications of the Triple Integral

Some applications of the triple integral are collected in **Table 8.11**, p. 533. The elementary areas corresponding to different coordinates are given in **Table 8.8**, p. 527. The elementary volumes corre-

sponding to different coordinates are given in **Table 8.10**.

Table 8.10 Elementary volumes

Coordinates	Elementary Volume
Cartesian coordinates x, y, z	$dV = dx\,dy\,dz$
Cylindrical coordinates $\rho\,, \varphi\,, z$	$dV = \rho\,d\rho\,d\varphi\,dz$
Spherical coordinates $r\,, \vartheta\,, \varphi$	$dV = r^2 \sin\vartheta\,dr\,d\vartheta\,d\varphi$
Arbitrary curvilinear coordinates u, v, w	$dV = \lvert D\rvert\,du\,dv\,dw$ (D Jacobian determinant)

8.5 Surface Integrals

There are distinguished surface integrals of the first type, of the second type, and of general type, analogously to the three different line integrals (see 8.3, p. 515).

8.5.1 Surface Integral of the First Type

The *surface integral* or *integral over a surface in space* is the generalization of the double integral, similarly as the line integral of the first type (see 8.3.1, p. 516) is a generalization of the ordinary integral.

8.5.1.1 Notion of the Surface Integral of the First Type

1. Definition

The *surface integral of the first type* of a function $u = f(x, y, z)$ of three variables defined in a connected domain is the integral

$$\int\limits_S f(x, y, z)\,dS, \tag{8.148a}$$

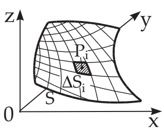


Figure 8.44

over a region S of a surface. The numerical value of the surface integral of the first kind is defined in the following way (see **Fig. 8.44**):

1. Decomposition of the region S in an arbitrary way into n elementary regions ΔS_i .
2. Choosing an arbitrary point $P_i(x_i, y_i, z_i)$ inside or on the boundary of each elementary region ΔS_i .
3. Multiplication of the value $f(x_i, y_i, z_i)$ of the function at this point by the area ΔS_i of the corresponding elementary region.
4. Summation of the products $f(x_i, y_i, z_i)\Delta S_i$.
5. Determination of the limit of the sum

$$\sum_{i=1}^{n} f(x_i, y_i, z_i)\,\Delta S_i \tag{8.148b}$$

as the diameter of each elementary region tends to zero, so ΔS_i tends to zero, hence, their number n tends to ∞ (see 8.4.1.1, **1.**, p. 524).
If this limit exists and is independent of the particular decomposition of the region S into elementary regions and also of the choice of the points $P_i(x_i, y_i, z_i)$, then it is called the *surface integral of the first type* of the function $u = f(x, y, z)$ over the region S, and one writes:

$$\int\limits_S f(x, y, z)\,dS = \lim_{\substack{\Delta S_i \to 0 \\ n\to\infty}} \sum_{i=1}^{n} f(x_i, y_i, z_i)\,\Delta S_i. \tag{8.148c}$$

2. Existence Theorem

If the function $f(x, y, z)$ is continuous on the domain, and the functions defining the surface have continuous derivatives here, the surface integral of the first type exists.

Table 8.11 Applications of the Triple Integral

General Formula	Cartesian Coordinates	Cylindrical Coordinates	Spherical Coordinates
1. Volume of a solid			
$V = \int_V dV =$	$\iiint dz\,dy\,dx$	$\iiint \rho\,dz\,d\rho\,d\varphi$	$\iiint r^2 \sin\vartheta\,dr\,d\vartheta\,d\varphi$
2. Axial moment of inertia of a solid with respect to the z-axis			
$I_z = \int_V \rho^2\,V =$	$\iiint (x^2+y^2)\,dz\,dy\,dx$	$\iiint \rho^3\,dz\,d\rho\,d\varphi$	$\iiint r^4 \sin^3\vartheta\,dr\,d\vartheta\,d\varphi$
3. Mass of a solid with the density function ϱ			
$M = \int_V \varrho\,dV =$	$\iiint \varrho\,dz\,dy\,dx$	$\iiint \varrho\rho\,dz\,d\rho\,d\varphi$	$\iiint \varrho r^2 \sin\vartheta\,dr\,d\vartheta\,d\varphi$
4. Coordinates of the center of a homogeneous solid			
$x_C = \dfrac{\int_V x\,dV}{V} =$	$\dfrac{\iiint x\,dz\,dy\,dx}{\iiint dz\,dy\,dx}$	$\dfrac{\iiint \rho^2\cos\varphi\,d\rho\,d\varphi\,dz}{\iiint \rho\,d\rho\,d\varphi\,dz}$	$\dfrac{\iiint r^3\sin^2\vartheta\cos\varphi\,dr\,d\vartheta\,d\varphi}{\iiint r^2\sin\vartheta\,dr\,d\vartheta\,d\varphi}$
$y_C = \dfrac{\int_V y\,dV}{V} =$	$\dfrac{\iiint y\,dz\,dy\,dx}{\iiint dz\,dy\,dx}$	$\dfrac{\iiint \rho^2\sin\varphi\,d\rho\,d\varphi\,dz}{\iiint \rho\,d\rho\,d\varphi\,dz}$	$\dfrac{\iiint r^3\sin^2\vartheta\sin\varphi\,dr\,d\vartheta\,d\varphi}{\iiint r^2\sin\vartheta\,dr\,d\vartheta\,d\varphi}$
$z_C = \dfrac{\int_V z\,dV}{V} =$	$\dfrac{\iiint z\,dz\,dy\,dx}{\iiint dz\,dy\,dx}$	$\dfrac{\iiint \rho z\,d\rho\,d\varphi\,dz}{\iiint \rho\,d\rho\,d\varphi\,dz}$	$\dfrac{\iiint r^3\sin\vartheta\cos\vartheta\,dr\,d\vartheta\,d\varphi}{\iiint r^2\sin\vartheta\,dr\,d\vartheta\,d\varphi}$

8.5.1.2 Evaluation of the Surface Integral of the First Type

The evaluation of the surface integral of the first type is reduced to the evaluation of a double integral over a planar domain (see 8.4.1, p. 524).

1. Explicit Representation of the Surface

If the surface S is given by the equation

$$z = z(x, y) \tag{8.149}$$

in explicit form, then

$$\int_S f(x,y,z)\,dS = \iint_{S'} f[x,y,z(x,y)]\sqrt{1+p^2+q^2}\,dx\,dy, \tag{8.150a}$$

is valid, where S' is the projection of S onto the x, y plane and p and q are the partial derivatives $p = \dfrac{\partial z}{\partial x}$, $q = \dfrac{\partial z}{\partial y}$. Here one assumes that to every point of the surface S there corresponds a unique point in S' in the x, y plane, i.e., the points of the surface are defined uniquely by their coordinates. If it does not hold, one decomposes S into several parts each of which satisfies the condition. Then the integral on the total surface can be calculated as the algebraic sum of the integrals over these parts of S.

The equation (8.150a) can be written in the form

$$\int_S f(x,y,z)\,dS = \iint_{S_{xy}} f[x,y,z(x,y)]\frac{d\,S_{xy}}{\cos\gamma}\,, \tag{8.150b}$$

since the equation of the surface normal of (8.149) has the form $\dfrac{X-x}{p} = \dfrac{Y-y}{q} = \dfrac{Z-z}{-1}$ (see **Table 3.29**, p. 264), since for the angle between the direction of the normal and the z-axis, $\cos\gamma = \dfrac{1}{\sqrt{1+p^2+q^2}}$ holds. In evaluating a surface integral of the first type, this angle γ is always considered as an acute angle, so $\cos\gamma > 0$ always holds.

2. Parametric Representation of the Surface

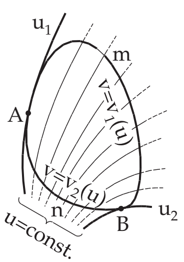


Figure 8.45

If the surface S is given in parametric form by the equations

$$x = x(u,v), \qquad y = y(u,v), \qquad z = z(u,v), \tag{8.151a}$$

(Fig. 8.45), then

$$\int_S f(x,y,z)\,dS$$

$$= \iint_\Delta f[x(u,v),y(u,v),z(u,v)]\sqrt{EG-F^2}\,du\,dv, \tag{8.151b}$$

where the functions E, F, and G are the quantities given in 3.6.3.3, **1.**, p. 263. The *elementary region in parametric form* is

$$\sqrt{EG-F^2}\,du\,dv = dS, \tag{8.151c}$$

and Δ is the domain of the parameters u and v corresponding to the given surface region. The evaluation is performed by a repeated integration with respect to v and u:

$$\int_S \Phi(u,v)\,dS = \int_{u_1}^{u_2}\int_{v_1(u)}^{v_2(u)} \Phi(u,v)\sqrt{EG-F^2}\,dv\,du, \quad \Phi = f[x(u,v),y(u,v),z(u,v)]. \tag{8.151d}$$

Here u_1 and u_2 are coordinates of the extreme coordinate lines $u = \text{const}$ enclosing the region S **(Fig. 8.45)**, and $v = v_1(u)$ and $v = v_2(u)$ are the equations of the curves $\overset{\frown}{AmB}$ and $\overset{\frown}{AnB}$ of the boundary of S.

Remark: The formula (8.150a) is a special case of (8.151b) for

$$u = x, \quad v = y, \quad E = 1+p^2, \quad F = p\,q, \quad G = 1+q^2. \tag{8.152}$$

3. Elementary Regions of Curved Surfaces

The elementary regions of curved surfaces are given in **Table 8.12**.

Table 8.12 Elementary Regions of Curved Surfaces

Coordinates	**Elementary Region**
Cartesian coordinates $x, y, z = z(x, y)$	$dS = \sqrt{1 + (\frac{\partial z}{\partial x})^2 + (\frac{\partial z}{\partial y})^2}\, dx\, dy$
Cylindrical lateral surface, R (const. radius), coordinates φ, z	$dS = R\, d\varphi\, dz$
Spherical surface R (const. radius), coordinates $\vartheta\,, \varphi$	$dS = R^2 \sin\vartheta\, d\vartheta\, d\varphi$
Arbitrary curvilinear coordinates u, v (E, F, G see differential of arc, p. 263)	$dS = \sqrt{EG - F^2}\, du\, dv$

8.5.1.3 Applications of the Surface Integral of the First Type

1. Surface Area of a Curved Surface

$$S = \int_S dS. \tag{8.153}$$

2. Mass of an Inhomogeneous Curved Surface S

With the coordinate-dependent density $\varrho = f(x, y, z)$ it follows:

$$M_S = \int_S \varrho\, dS. \tag{8.154}$$

8.5.2 Surface Integral of the Second Type

The *surface integral of the second type*, also called an *integral over a projection*, is a generalization of the notion of double integral similarly to the surface integral of the first type.

8.5.2.1 Notion of the Surface Integral of the Second Type

1. Notion of an Oriented Surface

A surface usually has two sides, and one of them can be chosen arbitrarily as the exterior one. If the exterior side is fixed, it is called an *oriented surface*. Surfaces for which one can not define two sides are not discussed here (see [8.7]).

2. Projection of an Oriented Surface onto a Coordinate Plane

Projecting a bounded part S of an oriented surface onto a coordinate plane, e.g., onto the x,y plane, one can consider this projection $Pr_{xy}\, S$ as positive or negative in the following way **(Fig. 8.46)**:

a) If the x, y plane is looked at from the positive direction of the z-axis, and one sees the positive side of the surface S, where the exterior part is considered to be positive, then the projection $Pr_{xy}\, S$ has a positive sign, otherwise it has a negative sign **(Fig. 8.46 a,b)**. **b)** If one part of the surface shows its positive side and the other part its negative side, then the projection $Pr_{xy}\, S$ is regarded as the algebraic sum of the positive and negative projections **(Fig. 8.46c)**.

The **Fig. 8.46d** shows the projections $Pr_{xz}S$ and $Pr_{yz}S$ of a surface S; one of them is positive the other one is negative.

The projection of a closed oriented surface is equal to zero.

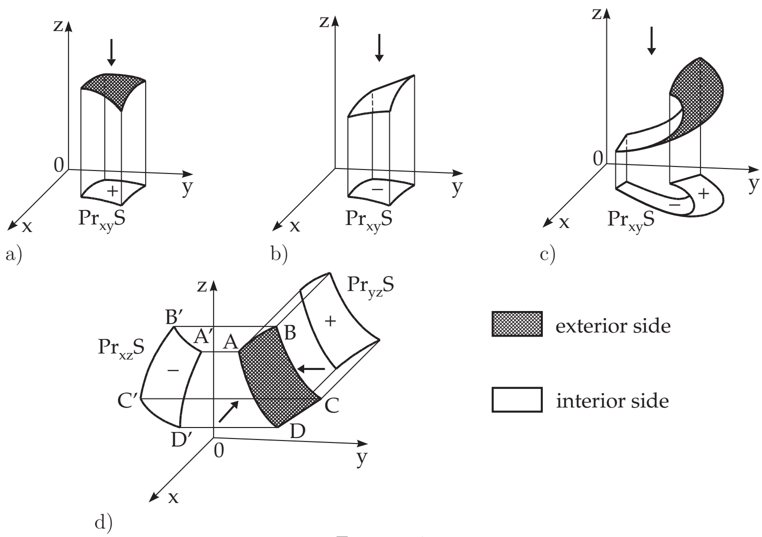


Figure 8.46

3. Definition of the Surface Integral of the Second Type over a Projection onto a Coordinate Plane

The *surface integral of the second type* of a function $f(x, y, z)$ of three variables defined in a connected domain is the integral

$$\int_S f(x, y, z)\, dx\, dy, \tag{8.155}$$

over the projection of an oriented surface S onto the x, y plane, where S is in the same domain where the function is defined, and if there is a one-to-one correspondence between the points of the surface and its projection. The numerical value of the integral is obtained in the same way as the surface integral of the first type except that in the third step the function value $f(x_i, y_i, z_i)$ is not multiplied by the elementary region ΔS_i, but by its projection $Pr_{xy}\Delta S_i$, oriented according to 8.5.2.1, **2.**, p. 535 on the x, y plane. Then holds:

$$\int_S f(x, y, z)\, dx\, dy = \lim_{\substack{\Delta S_i \to 0 \\ n \to \infty}} \sum_{i=1}^{n} f(x_i, y_i, z_i)\, Pr_{xy}\, \Delta S_i. \tag{8.156a}$$

Defining analogously the surface integrals of the second type over the projections of the oriented surface S onto the y, z plane and onto the z, x plane one gets:

$$\int_S f(x, y, z)\, dy\, dz = \lim_{\substack{\Delta S_i \to 0 \\ n \to \infty}} \sum_{i=1}^{n} f(x_i, y_i, z_i)\, Pr_{yz}\, \Delta S_i, \tag{8.156b}$$

$$\int_S f(x, y, z)\, dz\, dx = \lim_{\substack{\Delta S_i \to 0 \\ n \to \infty}} \sum_{i=1}^{n} f(x_i, y_i, z_i)\, Pr_{zx}\, \Delta S_i. \tag{8.156c}$$

4. Existence Theorem for the Surface Integral of the Second Type
The surface integral of the second type (8.156a,b,c) exists if the function $f(x, y, z)$ is continuous and the equations defining the surface are continuous and have continuous derivatives.

8.5.2.2 Evaluation of Surface Integrals of the Second Type

The principal method is to reduce it to the evaluation of double integrals.

1. Surface Given in Explicit Form
If the surface S is given by the equation

$$z = \varphi(x, y) \tag{8.157}$$

in explicit form, then the integral (8.156a) is calculated by the formula

$$\int_S f(x, y, z)\, dx\, dy = \int_{Pr_{xy}S} f[x, y, \varphi(x, y)]\, dS_{xy}, \tag{8.158a}$$

where $S_{xy} = Pr_{xy}S$. The surface integral of the function $f(x, y, z)$ over the projections of the surface S onto the other coordinate planes is calculated similarly:

$$\int_S f(x, y, z)\, dy\, dz = \int_{Pr_{yz}S} f[\psi(y, z), y, z]\, dS_{yz}, \tag{8.158b}$$

where one substitutes $x = \psi(y, z)$, the equation of the surface S solved for x, and $S_{yz} = Pr_{yz}S$.

$$\int_S f(x, y, z)\, dz\, dx = \int_{Pr_{zx}S} f[x, \chi(z, x), z]\, dS_{zx}, \tag{8.158c}$$

where one substitutes $y = \chi(z, x)$, the equation of the surface S solved for y, and $S_{zx} = Pr_{zx}S$. If the orientation of the surface is changed, i.e., if interchanging the exterior and interior sides, then the integral over the projection changes its sign.

2. Surface Given in Parametric Form
If the surface is given by the equations

$$x = x(u, v), \qquad y = y(u, v), \qquad z = z(u, v) \tag{8.159}$$

in parametric form, one calculates the integrals (8.156a,b,c) with help of the following formulas:

$$\int_S f(x, y, z)\, dx\, dy = \int_\Delta f[x(u, v), y(u, v), z(u, v)] \frac{D(x, y)}{D(u, v)}\, du\, dv, \tag{8.160a}$$

$$\int_S f(x, y, z)\, dy\, dz = \int_\Delta f[x(u, v), y(u, v), z(u, v)] \frac{D(y, z)}{D(u, v)}\, du\, dv, \tag{8.160b}$$

$$\int_S f(x, y, z)\, dz\, dx = \int_\Delta f[x(u, v), y(u, v), z(u, v)] \frac{D(z, x)}{D(u, v)}\, du\, dv. \tag{8.160c}$$

Here the expressions $\frac{D(x, y)}{D(u, v)}, \frac{D(y, z)}{D(u, v)}, \frac{D(z, x)}{D(u, v)}$ are the Jacobian determinants of pairs of functions x, y, z with respect to the variables u and v; Δ is the domain of u and v corresponding to the surface S.

8.5.3 Surface Integral in General Form

8.5.3.1 Notion of the Surface Integral in General Form

If $P(x, y, z),\ Q(x, y, z),\ R(x, y, z)$ are three functions of three variables defined in a connected domain and S is an oriented surface contained in this domain, the sum of the integrals of the second type taken

over the projections on the three coordinate planes is called the *surface integral in general form*:

$$\int_S (P\,dy\,dz + Q\,dz\,dx + R\,dx\,dy) = \int_S P\,dy\,dz + \int_S Q\,dz\,dx + \int_S R\,dx\,dy. \tag{8.161}$$

The formula reducing the surface integral to a double integral is:

$$\int_S (P\,dy\,dz + Q\,dz\,dx + R\,dx\,dy) = \int_\Delta \left[P\frac{D(y,z)}{D(u,v)} + Q\frac{D(z,x)}{D(u,v)} + R\frac{D(x,y)}{D(u,v)}\right] du\,dv, \tag{8.162}$$

where the quantities $\frac{D(x,y)}{D(u,v)}$, $\frac{D(y,z)}{D(u,v)}$, $\frac{D(z,x)}{D(u,v)}$, and Δ have the same meaning, as above.

Remark: The surface integral of vector-valued functions is discussed in the chapter about the theory of vector fields (see 13.3.2. p. 722).

8.5.3.2 Properties of the Surface Integrals

1. If the domain of integration, i.e., the surface S, is decomposed into two parts S_1 and S_2 (see **Fig. 8.47**), then

$$\int_S (P\,dy\,dz + Q\,dz\,dx + R\,dx\,dy) = \int_{S_1} (P\,dy\,dz + Q\,dz\,dx + R\,dx\,dy) + \int_{S_2} (P\,dy\,dz + Q\,dz\,dx + R\,dx\,dy). \tag{8.163}$$

2. If the orientation of the surface is reversed, i.e., the exterior and interior sides are interchanged, the integral changes its sign:

$$\int_{S^+} (P\,dy\,dz + Q\,dz\,dx + R\,dx\,dy) = -\int_{S^-} (P\,dy\,dz + Q\,dz\,dx + R\,dx\,dy), \tag{8.164}$$

where S^+ and S^- denote the same surface with different orientation.

3. A surface integral depends, in general, on the line bounding the surface region S as well as on the surface itself. Thus the integrals taken over two different non-closed surface regions S_1 and S_2 spanned by the same closed curve C are, in general, not equal **(Fig. 8.47)**:

$$\int_{S_1} (P\,dy\,dz + Q\,dz\,dx + R\,dx\,dy) \neq \int_{S_2} (P\,dy\,dz + Q\,dz\,dx + R\,dx\,dy). \tag{8.165}$$

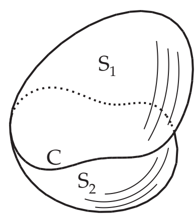


Figure 8.47

4. An application of the surface integral is the calculation of the volume V of a solid bounded by a closed surface S. The integral can be expressed and calculated in the form

$$V = \frac{1}{3}\int_S (x\,dy\,dz + y\,dz\,dx + z\,dx\,dy), \tag{8.166}$$

$$V = \int_S x\,dy\,dz \quad\text{or}\quad V = \int_S y\,dz\,dx \quad\text{or}\quad V = \int_S z\,dx\,dy \quad\text{or} \tag{8.167a}$$

$$V = \frac{1}{3}\int_S (x\,dy\,dz + y\,dz\,dx + z\,dx\,dy) \tag{8.167b}$$

where S is oriented so that its exterior side is positive.

■ Calculation of the volume V of a sphere with surface S according to $x^2 + y^2 + z^2 = R^2$: Using spherical coordinates $x = R\sin\vartheta\cos\varphi$, $y = R\sin\vartheta\sin\varphi$, $z = R\cos\vartheta$ $(0 \le \vartheta \le \pi,\ 0 \le \varphi \le 2\pi)$ and the Jacobian determinant as in (8.160a)

$$\frac{D(x,y)}{D(\vartheta,\varphi)} = \begin{vmatrix} x_\vartheta & x_\varphi \\ y_\vartheta & y_\varphi \end{vmatrix} = R^2\cos\vartheta\sin\vartheta \tag{8.168a}$$

follows from the third integral in (8.167a)

$$V = \int\limits_{\varphi=0}^{2\pi}\int\limits_{\vartheta=0}^{\pi} R^3\cos^2\vartheta\sin\vartheta\,d\vartheta\,d\varphi = 2\pi R^3\int\limits_{0}^{\pi}\cos^2\vartheta\sin\vartheta\,d\vartheta = \frac{4}{3}\pi R^3. \tag{8.168b}$$

9 Differential Equations

1. A Differential Equation is an equation, in which one or more variables, one or more functions of these variables, and also the derivatives of these functions with respect to these variables occur. The *order* of a differential equation is equal to the order of the highest occurring derivative.

2. Ordinary and Partial Differential Equations differ from each other in the number of their independent variables; in the first case there is only one, in the second case there are several.

■ **A:** $\left(\frac{dy}{dx}\right)^2 - xy^5\frac{dy}{dx} + \sin y = 0.$ ■ **B:** $x d^2y dx - dy(dx)^2 = e^y(dy)^3.$ ■ **C:** $\frac{\partial^2 z}{\partial x \partial y} = xyz\frac{\partial z}{\partial x}\frac{\partial z}{\partial y}.$

9.1 Ordinary Differential Equations

1. General Ordinary Differential Equation of Order *n*

in *implicit form* has the equation

$$F\left[x, y(x), y'(x), \ldots, y^{(n)}(x)\right] = 0. \tag{9.1}$$

If this equation is solved for $y^{(n)}(x)$, then it is the *explicit form* of an ordinary differential equation of order n.

2. Solution or Integral

of a differential equation is every function satisfying the equation in an interval $a \le x \le b$ which can be also infinite. A solution, which contains n arbitrary constants $c_1, c_2, \ldots, c_n$, is called the *general solution* or *general integral*. If the values of these constants are fixed, a *particular integral* or a *particular solution* is obtained. The value of these constants can be determined by n further conditions. If the values of y and its derivatives up to order $n-1$ are prescribed at one of the endpoints of the interval, then the problem is called an *initial value problem*. If there are given values at both endpoints of the interval, then the problem is called a *boundary value problem*.

■ The differential equation $-y'\sin x + y\cos x = 1$ has the general solution $y = \cos x + c\sin x$. For the condition $c = 0$ one gets the particular solution $y = \cos x$.

3. Initial Value Problem

If the n values $y(x_0), y'(x_0), \ldots, y^{(n-1)}(x_0)$ are given at x_0 for the solution $y = y(x)$ of an n-th order ordinary differential equation, then an *initial value problem* is given. The numbers are called the *initial values* or *initial conditions*. They form a system of n equations for the unknown constants $c_1, c_2, \ldots, c_n$ of the general solution of the n-th order ordinary differential equation.

■ The harmonic motion of a special elastic spring-mass system can be modeled by the initial value problem $y'' + y = 0$ with $y(0) = y_0, y'(0) = 0$. The solution is $y = y_0\cos x$.

4. Boundery Value Problem

If the solution of an ordinary differential equation and/or its derivatives are given at several points of its domain, then these values are called the *boundary conditions*. A differential equation with boundary conditions is called a *boundary value problem*.

■ The bending line of a bar with fixed endpoints and uniform load is described by the differential equation $y'' = x - x^2$ with the boundary conditions $y(0) = 0, y(1) = 0$ $(0 \le x \le 1)$. The solution is $y = \frac{x^3}{6} - \frac{x^4}{12} - \frac{x}{12}$.

9.1.1 First-Order Differential Equations

9.1.1.1 Existence Theorems, Direction Field

1. Existence of a Solution

In accordance with the Cauchy existence theorem the differential equation

$$y' = f(x, y) \tag{9.2}$$

I.N. Bronshtein et al., *Handbook of Mathematics*,
DOI 10.1007/978-3-662-46221-8_9

has at least one solution in a neighborhood of x_0 such that it takes the value y_0 at $x = x_0$ if the function $f(x, y)$ is continuous in a neighborhood G of the point (x_0, y_0). For example, G can be selected as the region given by $|x - x_0| < a$ and $|y - y_0| < b$ with some a and b.

2. Lipschitz Condition

The *Lipschitz condition* with respect to y is satisfied by $f(x, y)$ if

$$|f(x, y_1) - f(x, y_2)| \le N|y_1 - y_2| \tag{9.3}$$

holds for all (x, y_1) and (x, y_2) from G, where N is independent of x, y_1, and y_2. If this condition is satisfied, then the differential equation (9.2) has a unique solution through (x_0, y_0). The Lipschitz condition is obviously satisfied if $f(x, y)$ has a bounded partial derivative $\partial f/\partial y$ in this neighborhood. In 9.1.1.4, p. 546 there are examples in which the assumptions of the Cauchy existence theorem are not satisfied.

3. Direction Field

If the graph of a solution $y = \varphi(x)$ of the differential equation $y' = f(x, y)$ goes through the point $P(x, y)$, then the slope dy/dx of the tangent line of the graph at this point can be determined from the differential equation. So, at every point (x, y) the differential equation defines the slope of the tangent line of the solution passing through the considered point. The collection of these directions **(Fig. 9.1)** forms the *direction field*. An element of the direction field is a point together with the direction associated to it. Integration of a first-order differential equation geometrically means to connect the elements of a direction field into an *integral curve*, whose tangents have the same slopes at all points as the corresponding elements of the direction field.

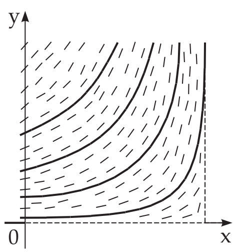


Figure 9.1

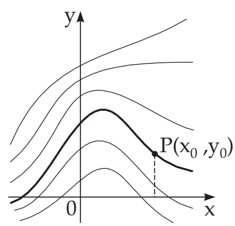


Figure 9.2

4. Vertical Directions

If a vertical direction can be found in a direction field, e.a., if the function $f(x, y)$ has a pole, then one can change the role of the independent and dependent variables and consider the differential equation

$$\frac{dx}{dy} = \frac{1}{f(x, y)} \tag{9.4}$$

as an equivalent equation to (9.2). In the region where the conditions of the existence theorems are fulfilled for the differential equations (9.2) or (9.4), there exists a unique integral curve **(Fig. 9.2)** through every point $P(x_0, y_0)$.

5. General Solution

The set of all integral curves of (9.2) can be characterized by one parameter and it can be given by the equation

$$F(x, y, C) = 0 \tag{9.5a}$$

of the corresponding one-parameter family of curves. The parameter C, an arbitrary constant, can be chosen freely and it is a necessary part of the general solution of every first-order differential equation.

A particular solution $y = \varphi(x)$, which satisfies the condition $y_0 = \varphi(x_0)$, can be obtained from the general solution (9.5a) if C is expressed from the equation

$$F(x_0, y_0, C) = 0. \tag{9.5b}$$

9.1.1.2 Important Solution Methods

1. Separation of Variables

If a differential equation can be transformed into the form

$$M(x)N(y)dx + P(x)Q(y)dy = 0, \tag{9.6a}$$

then it can be rewritten as

$$R(x)dx + S(y)dy = 0, \tag{9.6b}$$

where the variables x and y are separated into two terms. To get this form, equation (9.6a) is divided by $P(x)N(y)$. The general solution of (9.6a) is

$$\int \frac{M(x)}{P(x)}dx + \int \frac{Q(y)}{N(y)}dy = C. \tag{9.7}$$

If for some values $x = \overline{x}$ or $y = \overline{y}$, the functions $P(x)$ or $N(y)$ or both are equal to zero, then the constant functions $x = \overline{x}$ or/and $y = \overline{y}$ are also solutions of the differential equation. They are called *singular solutions*.

■ $xdy + ydx = 0$; $\int \frac{dy}{y} + \int \frac{dx}{x} = C$; $\ln|y| + \ln|x| = C = \ln|c|$; $yx = c$. If one allows also $c = 0$ in this final equation, then one has the singular solutions $y \equiv 0$ and $x \equiv 0$.

2. Homogeneous Equations

If $M(x, y)$ and $N(x, y)$ are homogeneous functions of the same order (see 2.18.2.6, **1.**, p. 122), then in the equation

$$M(x, y)dx + N(x, y)dy = 0 \tag{9.8}$$

the variables can be separated by substitution of $u = y/x$.

■ $x(x - y)y' + y^2 = 0$ with $y = u(x)x$, gives $(1 - u)u' + u/x = 0$, then by separation of the variables holds $\int \frac{(1-u)}{u}du = -\int \frac{1}{x}dx$. After integration: $\ln|x| + \ln|u - u| = C = \ln|c|$, $ux = ce^u$, $y = ce^{y/x}$. As can be seen in the preceding paragraph, Separation of Variables, the line $x = 0$ is also an integral curve.

3. Exact Differential Equations

An *exact differential equation* is an equation of the form

$$M(x, y)dx + N(x, y)dy = 0 \quad \text{or} \quad N(x, y)y' + M(x, y) = 0, \tag{9.9a}$$

if there exists a function $\Phi(x, y)$ of two variables such that

$$M(x, y)dx + N(x, y)dy \equiv d\Phi(x, y), \tag{9.9b}$$

i.e., if the left side of (9.9a) is the total differential of a function $\Phi(x, y)$ (see 6.2.2.1, p. 447). If functions $M(x, y)$ and $N(x, y)$ and their first-order partial derivatives are continuous on a connected domain G, then the equality

$$\frac{\partial M}{\partial y} = \frac{\partial N}{\partial x} \tag{9.9c}$$

is a necessary and sufficient condition for equation (9.9a) to be exact. In this case the general solution of (9.9a) is the function

$$\Phi(x, y) = C \quad (C = \text{const}), \tag{9.9d}$$

which can be calculated according to (8.3.4), 8.3.4.4, p. 522 as the integral

$$\Phi(x,y) = \int\limits_{x_0}^{x} M(\xi, y)\, d\xi + \int\limits_{y_0}^{y} N(x_0, \eta)\, d\eta, \tag{9.9e}$$

where x_0 and y_0 can be chosen arbitrarily from G.

■ Examples will be given later.

4. Integrating Factor

A function $\mu(x,y)$ is called an *integrating factor* or a *multiplier* if the equation

$$M dx + N dy = 0 \tag{9.10a}$$

multiplied by $\mu(x,y)$ becomes an exact differential equation. The integrating factor satisfies the differential equation

$$N\frac{\partial \ln \mu}{\partial x} - M\frac{\partial \ln \mu}{\partial y} = \frac{\partial M}{\partial y} - \frac{\partial N}{\partial x}. \tag{9.10b}$$

Every particular solution μ of this equation is an integrating factor. To give a general solution of this partial differential equation is much more complicated than to solve the original equation, so usually one is looking for the solution $\mu(x,y)$ in a special form, e.g., $\mu(x), \mu(y), \mu(xy)$ or $\mu(x^2+y^2)$.

■ To solve the differential equation $(x^2+y)\,dx - x\,dy = 0$, the equation for the integrating factor is $-x\frac{\partial \ln \mu}{\partial x} - (x^2+y)\frac{\partial \ln \mu}{\partial y} = 2$. An integrating factor which is independent of y must satisfy $x\frac{\partial \ln \mu}{\partial x} = -2$, so $\mu = \frac{1}{x^2}$. Multiplication of the given differential equation by μ yields $\left(1+\frac{y}{x^2}\right)dx - \frac{1}{x}dy = 0$. The general solution according to (9.9e) with the selection of $x_0 = 1, y_0 = 0$ is then:

$\Phi(x,y) \equiv \int_1^x \left(1+\frac{y}{\xi^2}\right) d\xi - \int_0^y d\eta = C \quad \text{or} \quad x - \frac{y}{x} = C_1.$

5. First-Order Linear Differential Equations

A *first-order linear differential equation* has the form

$$y' + P(x)y = Q(x), \tag{9.11a}$$

where the unknown function and its derivative occur only in first degree, and $P(x)$ and $Q(x)$ are given functions. If $P(x)$ and $Q(x)$ are continuous functions on a finite, closed interval, then the differential equation satisfies the conditions of the *Picard-Lindelöf theorem* (see 12.2.2.4,**4.**, p. 668) in this region. An integrating factor is here

$$\mu = \exp\left(\int P\, dx\right), \tag{9.11b}$$

the general solution is

$$y = \exp\left(-\int P\, dx\right)\left[\int Q \exp\left(\int P\, dx\right) dx + C\right]. \tag{9.11c}$$

Replacing the indefinite integrals by definite ones with lower bound x_0 and upper bound x in this formula, then for the solution $y(x_0) = C$ (see 8.2.1.2, **1.**, p. 495). If y_1 is any particular solution of the differential equation, then the general solution of the differential equation is given by the formula

$$y = y_1 + C\exp\left(-\int P\, dx\right). \tag{9.11d}$$

If $y_1(x)$ and $y_2(x)$ are two linearly independent particular solutions (see 9.1.2.3, **2.**, p. 553), then one can get the general solution without any integration as

$$y = y_1 + C(y_2 - y_1). \tag{9.11e}$$

■ To solve the differential equation $y' - y\tan x = \cos x$ with the initial condition $x_0 = 0$, $y_0 = 0$. Calculating $\exp\left(-\int_0^x \tan x\,dx\right) = \cos x$ one gets the solution according to (9.11c):

$$y = \frac{1}{\cos x}\int_0^x \cos^2 x\,dx = \frac{1}{\cos x}\left[\frac{\sin x\cos x + x}{2}\right] = \frac{\sin x}{2} + \frac{x}{2\cos x}.$$

6. Bernoulli Differential Equations

The *Bernoulli differential equation* is an equation of the form

$$y' + P(x)y = Q(x)y^n \quad (n \neq 0, n \neq 1), \tag{9.12}$$

which can be reduced to a linear differential equation if it is divided by y^n and the new variable $z = y^{-n+1}$ is introduced.

■ Solution of the differential equation $y' - \dfrac{4y}{x} = x\sqrt{y}$. Since $n = 1/2$, dividing by $\sqrt{y}$ and introducing the new variable $z = \sqrt{y}$ leads to the equation $\dfrac{dz}{dx} - \dfrac{2z}{x} = \dfrac{x}{2}$. By using the formulas for the solution of a linear differential equation there is $\exp(\int P\,dx) = \dfrac{1}{x^2}$ and $z = x^2\left[\int \dfrac{x}{2}\dfrac{1}{x^2}\,dx + C\right] = x^2\left[\dfrac{1}{2}\ln|x| + C\right]$.

So, finally, $y = x^4\left(\dfrac{1}{2}\ln|x| + C\right)^2$.

7. Riccati Differential Equations

The *Riccati differential equation*

$$y' = P(x)y^2 + Q(x)y + R(x), \tag{9.13a}$$

usually cannot be solved by elementary integration, i.e., not by using a final number of successive elementary integrations. However it is possible to transform it by suitable substitutions into differential equations for which solutions often can be found.

Method 1: By the substitution

$$y = \frac{u(x)}{P(x)} + \beta(x) \tag{9.13b}$$

the Riccati differential equation can be transformed into the *normal form*

$$\frac{du}{dx} = u^2 + R_0(x) \quad (9.13c) \qquad \text{with} \quad R_0(x) = P^2\beta^2 + QP\beta + PR - P\beta'. \tag{9.13d}$$

Therefore $\beta(x)$ is determined so that terms with the factor $u(x)$ disappear.
If a particular solution $u_1(x)$ of (9.13c) is known, which can be found, e.g., by a suitable approach, then by the help of the substitution

$$u = \frac{1}{z(x)} + u_1(x) \tag{9.13e}$$

(9.13c) is to be transformed into the linear differential equation for $z(x)$:

$$z' + 2u_1(x)z - 1 = 0. \tag{9.13f}$$

From the solution of (9.13f) the solution of (9.13a) is obtained by using (9.13e) and (9.13b).

Method 2: By the substitution

$$y = -\frac{v'}{P(x)v(x)} \tag{9.13g}$$

(9.13a) is transformed into a linear differential equation of second order (see 9.1.2.6,**1.**, p. 560):

$$Pv'' - (P' + PQ)v' + P^2Rv = 0. \tag{9.13h}$$

■ To solve the differential equation $y' + y^2 + \frac{1}{x}y - \frac{4}{x^2} = 0$, i.e. for $P = -1, Q = -\frac{1}{x}, R = \frac{4}{x^2}$.

Method 1: One gets $\beta(x) = -\frac{1}{2x}$ and with the help of $y = -u(x) - \frac{1}{2x}$ one gets the normal form $u' = u^2 - \frac{15}{4x^2}$. Particular solutions of the normal form can be got, e.g., with the approach $u = \frac{a}{x}$: $u_1(x) = \frac{3}{2x}$, $u_2(x) = -\frac{5}{2x}$. After substituting $u = \frac{1}{z(x)} + \frac{3}{2x}$ the differential equation $z' + \frac{3}{x}z + 1 = 0$ follows with the solution $z(x) = -\frac{x}{4} + \frac{K}{x^3} = \frac{4K - x^4}{4x^3}$ (K const). The inverse transformation gives $y = \frac{2x^4 + 2C}{x^5 - Cx}$ $(C = 4K)$.

Method 2: According to (9.13h) the Euler differential equation $x^2v'' + xv' - 4v = 0$ is obtained with the general solution $v(x) = C_1x^2 + C_2\frac{1}{x^2}$ (see ■ concerning the Euler differential equation, p. 557). One of the constants C_1 and C_2 can be chosen freely, e.g. $C_2 = -1$ then from (9.13h) follows $y = \frac{2x^4 + 2C_1}{x^5 - C_1x}$.

9.1.1.3 Implicit Differential Equations

1. Solution in Parametric Form

Given a differential equation in implicit form

$$F(x, y, y') = 0. \tag{9.14}$$

There are n integral curves passing through a point $P(x_0, y_0)$ if the following conditions hold:

a) The equation $F(x_0, y_0, p) = 0$ $(p = dy/dx)$ has n real roots $p_1, \ldots, p_n$ at the point $P(x_0, y_0)$.

b) The function $F(x, y, p)$ and its first partial derivatives are continuous at $x = x_0$, $y = y_0$, $p = p_i$; furthermore $\partial F/\partial p \neq 0$.

If the original equation can be solved with respect to y', then it yields n equations of the explicit forms discussed above. Solving these equations one gets n families of integral curves. If the equation can be written in the form $x = \varphi(y, y')$ or $y = \psi(x, y')$, then putting $y' = p$ and considering p as an auxiliary variable, after differentiation with respect to y or x one obtains an equation for dp/dy or dp/dx which is solved with respect to the derivative. A solution of this equation together with the original equation (9.14) determines a desired solution in parametric form.

■ To get the solution of the differential equation $x = yy' + y'^2$, one substitutes $y' = p$ and gets $x = py + p^2$. Differentiation with respect to y and substituting $\frac{dx}{dy} = \frac{1}{p}$ results in $\frac{1}{p} = p + (y + 2p)\frac{dp}{dy}$ or $\frac{dy}{dp} - \frac{py}{1-p^2} = \frac{2p^2}{1-p^2}$. Solving this equation for y one obtains $y = -p + \frac{C + \arcsin p}{\sqrt{1-p^2}}$ (C const). Substitution into the initial equation gives the solution for x in parametric form.

2. Lagrange Differential Equation

The *Lagrange differential equation* is the equation

$$a(y')x + b(y')y + c(y') = 0. \tag{9.15a}$$

The solution can be determined by the method given above. If for $p = p_0$ holds

$$a(p) + b(p)p = 0, \tag{9.15b}$$

then

$$a(p_0)x + b(p_0)y + c(p_0) = 0 \tag{9.15c}$$

is a singular solution of (9.15a).

3. Clairaut Differential Equation

The *Clairaut differential equation* is the special case of the Lagrange differential equation if

$$a(p) + b(p)p \equiv 0, \tag{9.16a}$$

and so it can be transformed into the form

$$y = y'x + f(y'). \tag{9.16b}$$

The general solution is

$$y = Cx + f(C). \tag{9.16c}$$

Besides the general solution, the Clairaut differential equation also has a singular solution, which can be obtained by eliminating the constant C from the equations

$$y = Cx + f(C) \quad (9.16\text{d}) \qquad \text{and} \quad 0 = x + f'(C), \tag{9.16e}$$

The second equation can be obtained by differentiating the first one with respect to C. Geometrically, the singular solution is the envelope (see 3.6.1.7, p. 255) of the solution family of lines **(Fig. 9.3)**.

■ Solution of the differential equation $y = xy' + y'^2$. The general solution is $y = Cx + C^2$. The singular solution one gets with the help of the equation $x + 2C = 0$ to eliminate C, and hence $x^2 + 4y = 0$. **Fig. 9.3** shows this case.

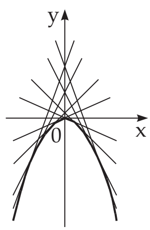


Figure 9.3

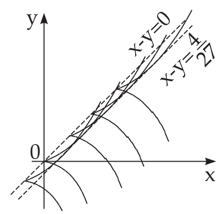


Figure 9.4

9.1.1.4 Singular Integrals and Singular Points

1. Singular element

An element (x_0, y_0, y_0') is called a *singular element* of the differential equation, if in addition to the differential equation

$$F(x, y, y') = 0 \tag{9.17a}$$

it also satisfies the equation

$$\frac{\partial F}{\partial y'} = 0. \tag{9.17b}$$

2. Singular Integral

An integral curve from singular elements is called a *singular integral curve*; the equation

$$\varphi(x, y) = 0 \tag{9.17c}$$

of a singular integral curve is called a *singular integral*. The envelopes of the integral curves are singular integral curves **(Fig. 9.3)**; they consist of the singular elements.

The uniqueness of the solution (see 9.1.1.1, **1.**, p. 540) usually fails at the points of a singular integral curve.

3. Determination of Singular Integrals

Usually one cannot obtain singular integrals for any values of the arbitrary constants of the general solution. To determine the singular solution of a differential equation (9.17a) with $p = y'$ one has to

introduce the equation

$$\frac{\partial F}{\partial p} = 0 \tag{9.17d}$$

and to eliminate p. If the obtained relation is a solution of the given differential equation, then it is a singular solution. The equation of this solution should be transformed into a form which does not contain multiple-valued functions, in particular no radicals where the complex values should also be considered.

Radicals are expressions obtained by nesting algebraic equations (see 2.2.1, p. 62). If the equation of the family of integral curves is known, i.e., the general solution of the given differential equation is known, then one can determine the envelope of the family of curves, the singular integral, with the methods of differential geometry (see 3.6.1.7, p. 255).

■ Solution of the differential equation $x - y - \frac{4}{9}y'^2 + \frac{8}{27}y'^3 = 0$. Substituting $y' = p$, the calculation of the additional equation with (9.17d) yields $-\frac{8}{9}p + \frac{8}{9}p^2 = 0$. Elimination of p results in equation a) $x - y = 0$ and b) $x - y = \frac{4}{27}$, where a) is not a solution, b) is a solution, a special case of the general solution $(y - C)^2 = (x - C)^3$. The integral curves of a) and b) are shown in **Fig. 9.4**.

4. Singular Points of a Differential Equation

Singular points of a differential equation are the points where the right side of the differential equation

$$y' = f(x, y) \tag{9.18a}$$

is not defined. This is the case, e.g., in the differential equations of the following forms:

1. Differential Equation with a Fraction of Linear Functions

$$\frac{dy}{dx} = \frac{ax + by}{cx + ey} \qquad (ae - bc \neq 0) \tag{9.18b}$$

has an *isolated singular point* at $(0, 0)$, since the assumptions of the existence theorem are fulfilled almost at every point arbitrarily close to $(0, 0)$ but not at this point itself. The conditions are not fulfilled at the points where $cx + ey = 0$. One can force the fulfillment of the conditions at these points exchanging the role of the variables and considering the equation

$$\frac{dx}{dy} = \frac{cx + ey}{ax + by}\,. \tag{9.18c}$$

The behavior of the integral curve in the neighborhood of a singular point depends on the roots of the *characteristic equation*

$$\lambda^2 - (b + c)\lambda + bc - ae = 0. \tag{9.18d}$$

The following cases can be distinguished:

Case 1: If the roots are real and they have the same sign, then the singular point is a *branch point.* The integral curves in a neighborhood of the singular point pass through it and if the roots of the characteristic equation do not coincide, they have a common tangent except for one. If the roots coincide, then either all integral curves have the same tangent, or there is a unique integral curve passing through the singular point in each direction.

■ **A:** For the differential equation $\frac{dy}{dx} = \frac{2y}{x}$ the characteristic equation is $\lambda^2 - 3\lambda + 2 = 0,\quad \lambda_1 = 2,$ $\lambda_2 = 1$. The integral curves have the equation $y = C\,x^2$ **(Fig. 9.5)**. The general solution also contains the line $x = 0$ considering the form $x^2 = C_1\,y$.

■ **B:** The characteristic equation for $\frac{dy}{dx} = \frac{x + y}{x}$ is $\lambda^2 - 2\lambda + 1 = 0$, $\lambda_1 = \lambda_2 = 1$. The integral curves

are $y = x \ln |x| + Cx$ **(Fig. 9.6)**. The singular point is a so-called *node*.

■ **C:** The characteristic equation for $\frac{dy}{dx} = \frac{y}{x}$ is $\lambda^2 - 2\lambda + 1 = 0$, $\lambda_1 = \lambda_2 = 1$. The integral curves are $y = C\,x$ **(Fig. 9.7)**. The singular point is a so-called *ray point*.

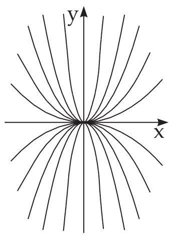


Figure 9.5

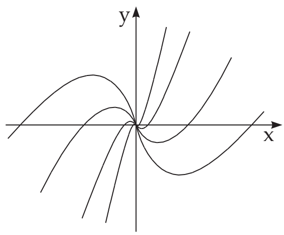


Figure 9.6

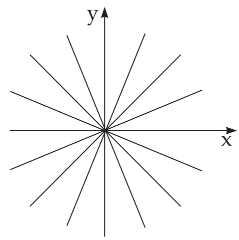


Figure 9.7

Case 2: If the roots are real and they have different signs, the singular point is a *saddle point*, and two of the integral curves pass through it.

■ **D:** The characteristic equation for $\frac{dy}{dx} = -\frac{y}{x}$ is $\lambda^2 - 1 = 0$, $\lambda_1 = +1$, $\lambda_2 = -1$. The integral curves are $x\,y = C$ **(Fig. 9.8)**. For $C = 0$ the particular solutions $x = 0$, $y = 0$ hold.

Case 3: If the roots are conjugate complex numbers with a non-zero real part ($\mathrm{Re}(\lambda) \neq 0$), then the singular point is a *spiral point* which is also called a *focal point*, and the integral curves wind about this singular point.

■ **E:** The characteristic equation for $\frac{dy}{dx} = \frac{x+y}{x-y}$ is $\lambda^2 - 2\lambda + 2 = 0$, $\lambda_1 = 1 + \mathrm{i}$, $\lambda_2 = 1 - \mathrm{i}$. The integral curves in polar coordinates are $r = C\,e^{\varphi}$ **(Fig. 9.9)**.

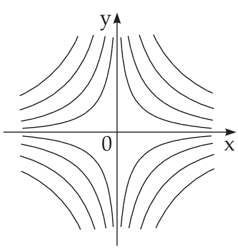


Figure 9.8

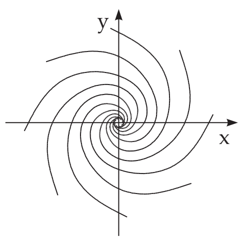


Figure 9.9

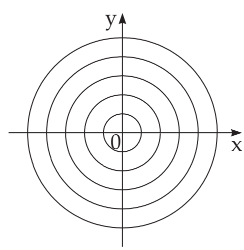


Figure 9.10

Case 4: If the roots are pure imaginary numbers, then the singular point is a *central point*, or *center*, which is surrounded by the closed integral curves.

■ **F:** The characteristic equation for $\frac{dy}{dx} = -\frac{x}{y}$ is $\lambda^2 + 1 = 0$, $\lambda_1 = \mathrm{i}$, $\lambda_2 = -\mathrm{i}$. The integral curves are $x^2 + y^2 = C$ **(Fig. 9.10)**.

2. Differential Equation with the Ratio of Two Arbitrary Functions

$$\frac{dy}{dx} = \frac{P(x,y)}{Q(x,y)} \tag{9.19a}$$

has the singular points for the values of the variables where

$$P(x,y) = Q(x,y) = 0. \tag{9.19b}$$

If P and Q are continuous functions and they have continuous partial derivatives, (9.19a) can be written in the form

$$\frac{dy}{dx} = \frac{a(x-x_0) + b(y-y_0) + P_1(x,y)}{c(x-x_0) + e(y-y_0) + Q_1(x,y)}. \tag{9.19c}$$

Here x_0 and y_0 are the coordinates of the singular point and $P_1(x,y)$ and $Q_1(x,y)$ are infinitesimals of a higher order than the distance of the point (x,y) from the singular point (x_0,y_0). With these assumptions the type of a singular point of the given differential equation is the same as that of the *approximate equation* obtained by omitting the terms P_1 and Q_1, with the following exceptions:

a) If the singular point of the approximate equation is a center, the singular point of the original equation is either a center or a focal point.

b) If $a\,e - b\,c = 0$, i.e., $\frac{a}{c} = \frac{b}{e}$ or $a = c = 0$ or $a = b = 0$, then the type of singular point should be determined by examining the terms of higher order.

9.1.1.5 Approximation Methods for Solution of First-Order Differential Equations

1. Successive Approximation Method of Picard

The integration of the differential equation

$$y' = f(x,y) \tag{9.20a}$$

with the initial condition $y = y_0$ for $x = x_0$ results in the fixed-point problem

$$y = y_0 + \int_{x_0}^{x} f(x,y)\,dx. \tag{9.20b}$$

Substituting another function $y_1(x)$ instead of y into the right-hand side of (9.20b), then the result will be a new function $y_2(x)$, which is different from $y_1(x)$, if $y_1(x)$ is not already a solution of (9.20a). Substituting $y_2(x)$ instead of y into the right-hand side of (9.20b) gives a function $y_3(x)$. If the conditions of the existence theorem are fulfilled (see 9.1.1.1, **1.**, p. 540), the sequence of functions $y_1, y_2, y_3, \ldots$ converges to the desired solution in a certain interval containing the point x_0.

This *Picard method of successive approximation* is an *iteration method* (see 19.1.1, p. 949).

■ Solve the differential equation $y' = e^x - y^2$ with initial values $x_0 = 0$, $y_0 = 0$. Rewriting the equation in integral form and using the successive approximation method with an initial approximation $y_0(x) \equiv 0$ gives: $y_1 = \int_0^x e^x\,dx = e^x - 1$, $\quad y_2 = \int_0^x \left[e^x - (e^x-1)^2\right] dx = 3e^x - \frac{1}{2}e^{2x} - x - \frac{5}{2}$, etc.

2. Solution by Series Expansion

The Taylor series expansion of the solution of a differential equation (see 7.3.3.3, **1.**, p. 471) can be given in the form

$$y = y_0 + (x-x_0){y_0}' + \frac{(x-x_0)^2}{2}{y_0}'' + \cdots + \frac{(x-x_0)^n}{n!}{y_0}^{(n)} + \cdots \tag{9.21}$$

if the values ${y_0}'$, ${y_0}''$, ..., ${y_0}^{(n)}, \ldots$ of all derivatives of the solution function are known at the initial value x_0 of the independent variable. The values of the derivatives can be determined by successively differentiating the original equation and substituting the initial conditions. If the differential equation can be differentiated infinitely many times, the obtained series will be convergent in a certain neighborhood of the initial value of the independent variable. This method can be used also for n-th order differential equations.

Remark: The above result is the Taylor series of the function, which may not represent the function

itself (see 7.3.3.3, **1.**, p. 471).

It is often useful to substitute the solution by an infinite series with unknown coefficients, and to determine them by comparing coefficients.

■ **A:** To solve the differential equation $y' = e^x - y^2$, $x_0 = 0$, $y_0 = 0$ one can consider the series $y = a_1x + a_2x^2 + a_3x^3 + \cdots + a_nx^n + \cdots$. Substituting this into the equation considering the formula (7.88), p. 470 for the square of the series gives

$$a_1 + 2a_2x + 3a_3x^2 + \cdots + [{a_1}^2x^2 + 2a_1a_2x^3 + ({a_2}^2 + 2a_1a_3)x^4 + \cdots] = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \cdots.$$

Comparing coefficients gives: $a_1 = 1$, $2a_2 = 1$, $3a_3 + {a_1}^2 = \frac{1}{2}$, $4a_4 + 2a_1a_2 = \frac{1}{6}$, etc. Solving these equations successively and substituting the coefficient values into the series representation yields $y = x + \frac{x^2}{2} - \frac{x^3}{6} - \frac{5}{24}x^4 + \cdots$.

■ **B:** The same differential equation with the same initial conditions can also be solved in the following way: Substituting $x = 0$ into the equation, gives ${y_0}' = 1$ and successive differentiation yields $y'' = e^x - 2yy'$, ${y_0}'' = 1$, $y''' = e^x - 2y'^2 - 2yy''$, ${y_0}''' = -1$, $y^{(4)} = e^x - 6y'y'' - 2yy'''$, ${y_0}^{(4)} = -5$, *etc.* From the Taylor theorem (see 7.3.3.3, **1.**, p. 471) follows the solution $y = x + \frac{x^2}{2!} - \frac{x^3}{3!} - \frac{5x^4}{4!} + \cdots$.

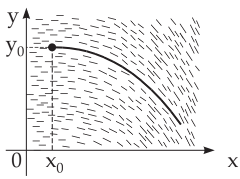


Figure 9.11

3. Graphical Solution of Differential Equations

The graphical integration of a differential equation is a method, which is based on the direction field (see 9.1.1.1, **3.**, p. 541). The integral curve in **Fig. 9.11** is represented by a broken line which starts at the given initial point and is composed of short line segments. The directions of the line segments are always the same as the direction of the direction field at the starting point of the line segment. This is also the endpoint of the previous line segment.

4. Numerical Solution of Differential Equations

The numerical solutions of differential equations will be discussed in detail in 19.4, p. 969. Numerical methods are used to determine a solution of a differential equation, if the equation $y' = f(x, y)$ does not belong to the special cases discussed above whose analytic solutions are known, or if the function $f(x, y)$ is too complicated. This can happen if $f(x, y)$ is non-linear in y.

9.1.2 Differential Equations of Higher Order and Systems of Differential Equations

9.1.2.1 Basic Results

1. Existence of a Solution

1. Reduction to a System of Differential Equations Every explicit n-th order differential equation

$$y^{(n)} = f\left(x, y, y', \ldots, y^{(n-1)}\right) \tag{9.22a}$$

by introducing the new variables

$$y_1 = y', \quad y_2 = y'', \ldots, y_{n-1} = y^{(n-1)} \tag{9.22b}$$

can be reduced to a system of n first-order differential equations

$$\frac{dy}{dx} = y_1, \quad \frac{dy_1}{dx} = y_2, \ldots, \frac{dy_{n-1}}{dx} = f(x, y, y_1, \ldots, y_{n-1}). \tag{9.22c}$$

2. Existence of a System of Solutions The system of n differential equations

$$\frac{dy_i}{dx} = f_i(x, y_1, y_2, \dots, y_n) \qquad (i = 1, 2, \dots, n), \tag{9.23a}$$

which is more general than system (9.22c), has a unique system of solutions

$$y_i = y_i(x) \quad (i = 1, 2, \dots, n), \tag{9.23b}$$

which is defined in an interval $x_0 - h \le x \le x_0 + h$ and for $x = x_0$ takes the previously given initial values $y_i(x_0) = {y_i}^0$ $(i = 1, 2, \dots, n)$, if the functions $f_i(x, y_1, y_2, \dots, y_n)$ are continuous with respect to all variables and satisfy the following Lipschitz condition.

3. Lipschitz condition For the values x, y_i and $y_i + \Delta y_i$, which are in a certain neighborhood of the given initial values, the functions f_i satisfy the following inequalities:

$$\begin{aligned} &|f_i(x, y_1 + \Delta y_1, y_2 + \Delta y_2, \dots, y_n + \Delta y_n) - f_i(x, y_1, y_2, \dots, y_n)| \\ &\qquad \le K\,(|\Delta y_1| + |\Delta y_2| + \dots + |\Delta y_n|) \end{aligned} \tag{9.24a}$$

with a common constant K (see also 9.1.1.1, **2.**, p. 541).

This fact implies that if the function $f(x, y, y', \dots, y^{(n-1)})$ is continuous and satisfies the Lipschitz condition (9.24a), then the equation

$$y^{(n)} = f\left(x, y, y', \dots, y^{(n-1)}\right) \tag{9.24b}$$

has a unique solution with the initial values $y(x_0) = y_0$, $y'(x_0) = {y_0}'$, ..., $y^{(n-1)}(x_0) = {y_0}^{(n-1)}$, and it is $(n-1)$ times continuously differentiable.

2. General Solution

1. The general solution of the differential equation (9.24b) contains n independent arbitrary constants:

$$y = y(x, C_1, C_2, \dots, C_n). \tag{9.25a}$$

2. In the geometrical interpretation the equation (9.25a) defines a family of curves depending on n parameters. Every single one of these integral curves, i.e., the graph of the corresponding particular solution, can be obtained by a suitable choice of the constants $C_1, C_2, \dots, C_n$. If the solution has to satisfy the above initial conditions, then the values $C_1, C_2, \dots, C_n$ are determined from the following equations:

$$\begin{aligned} y(x_0, C_1, \dots, C_n) &= y_0, \\ \left[\frac{d}{dx} y(x, C_1, \dots, C_n)\right]_{x=x_0} &= {y_0}', \\ \dots\dots\dots\dots\dots\dots\dots\dots & \\ \left[\frac{d^{n-1}}{dx^{n-1}} y(x, C_1, \dots, C_n)\right]_{x=x_0} &= {y_0}^{(n-1)}. \end{aligned} \tag{9.25b}$$

If these equations are inconsistent for any initial values in a certain domain, then the solution is not general in this domain, i.e., the arbitrary constants cannot be chosen independently.

3. The general solution of system (9.23a) also contains n arbitrary constants. This general solution can be represented in two different ways: Either it is given in a form which is solved for the unknown functions

$$y_1 = F_1(x, C_1, \dots, C_n),\;\; y_2 = F_2(x, C_1, \dots, C_n), \dots, y_n = F_n(x, C_1, \dots, C_n) \tag{9.26a}$$

or in the form which is solved for the constants

$$\varphi_1(x, y_1, \dots, y_n) = C_1,\;\; \varphi_2(x, y_1, \dots, y_n) = C_2, \dots, \varphi_n(x, y_1, \dots, y_n) = C_n. \tag{9.26b}$$

In the case of (9.26b) each relation

$$\varphi_i(x, y_1, \dots, y_n) = C_i \tag{9.26c}$$

is a *first integral* of the system (9.23a). The first integral can be defined independently of the general solution as a relation (9.26c). That is, (9.26c) will be an identity replacing $y_1, y_2, \ldots, y_n$ by any particular solution of the given system and replacing the constant by the arbitrary constant C_i determined by this particular solution.
If any first integral is known in the form (9.26c), then the function $\varphi_i(x, y_1, \ldots, y_n)$ satisfies the partial different equation

$$\frac{\partial\varphi_i}{\partial x} + f_1(x, y_1, \ldots, y_n)\frac{\partial\varphi_i}{\partial y_1} + \cdots + f_n(x, y_1, \ldots, y_n)\frac{\partial\varphi_i}{\partial y_n} = 0. \quad (9.26d)$$

Conversely, each solution $\varphi_i(x, y_1, \ldots, y_n)$ of the partial differential equation (9.26d) defines a first integral of the system (9.23a) in the form (9.26c). The general solution of the system (9.23a) can be represented as a system of n first integrals of system (9.23a), if the corresponding functions $\varphi_i(x, y_1, \ldots, y_n)$ $(i = 1, 2, \ldots, n)$ are linearly independent (see 9.1.2.3, **2.**, p. 553).

9.1.2.2 Lowering the Order

One of the most important solution methods for n-th order differential equations

$$f\left(x, y, y', \ldots, y^{(n)}\right) = 0 \quad (9.27)$$

is the substitution of variables in order to obtain a simpler differential equation, especially one of lower order. Different cases can be distinguished.

1. $\boldsymbol{f = f(y, y', \ldots, y^{(n)})}$**, i.e.,** $\boldsymbol{x}$ **does not appear explicitly:**

$$f\left(y, y', \ldots, y^{(n)}\right) = 0. \quad (9.28a)$$

By substitution

$$\frac{dy}{dx} = p, \quad \frac{d^2y}{dx^2} = p\frac{dp}{dy}, \ldots \quad (9.28b)$$

the order of the differential equation can be reduced from n to $(n-1)$.

■ Reducing the order of the differential equation $yy'' - y'^2 = 0$ to one, with the substitution $y' = p, p\,dp/dy = y''$ it becomes a first-order differential equation $y\,p\,dp/dy - p^2 = 0$, and $y\,dp/dy - p = 0$ results in $p = C\,y = dy/dx$, $y = C_1 e^{Cx}$. Canceling p does not result in a loss of a solution, since $p = 0$ gives the solution $y = C_1$, which is included in the general solution with $C = 0$.

2. $\boldsymbol{f = f(x, y', \ldots, y^{(n)})}$**, i.e.,** $\boldsymbol{y}$ **does not appear explicitly:**

$$f\left(x, y', ..., y^{(n)}\right) = 0. \quad (9.29a)$$

The order of the differential equation can be reduced from n to $(n-1)$ by the substitution

$$y' = p. \quad (9.29b)$$

If the first k derivatives are missing in the initial equation, then a suitable substitution is

$$y^{(k+1)} = p. \quad (9.29c)$$

■ The order of the differential equation $y'' - xy''' + (y''')^3 = 0$ will be reduced by the substitution $y'' = p$, so one gets a Clairaut differential equation $p - x\dfrac{dp}{dx} + \left(\dfrac{dp}{dx}\right)^3 = 0$ whose general solution is $p = C_1\,x + {C_1}^3$. Therefore, $y = \dfrac{C_1x^3}{6} - \dfrac{{C_1}^3x^2}{2} + C_2x + C_3$. From the singular solution of the Clairaut differential equation $p = \dfrac{2\sqrt{3}}{9}x^{3/2}$ one gets the singular solution of the original equation: $y = \dfrac{8\sqrt{3}}{315}x^{7/2} + C_1x + C_2$.

3. $f\left(x, y, y', \dots, y^{(n)}\right)$ is a homogeneous function (see 2.18.2.4, **4.**, p. 122) **in $y, y', y'', \dots, y^{(n)}$:**

$$f\left(x, y, y', \dots, y^{(n)}\right) = 0. \tag{9.30a}$$

One can reduce the order by the substitution

$$z = \frac{y'}{y}, \quad \text{i.e.,} \quad y = e^{\int z\,dx}. \tag{9.30b}$$

■ Transforming the differential equation $yy'' - y'^2 = 0$ by the substitution $z = y'/y$, results in $\frac{dz}{dx} = \frac{yy'' - y'^2}{y^2} = 0$ so the order is reduced by one. One gets $z = C_1$, therefore, $\ln|y| = C_1 x + C_2$, or $y = Ce^{C_1 x}$ with $\ln|C| = C_2$.

4. $f = f(x, y, y', \dots, y^{(n)})$ is a function of only x:

$$y^{(n)} = f(x). \tag{9.31a}$$

One gets the general solution by n repeated integrations. It has the form

$$y = C_1 + C_2 x + C_3 x^2 + \cdots + C_n x^{n-1} + \psi(x) \tag{9.31b}$$

with

$$\psi(x) = \iint \cdots \int f(x)\,(dx)^n = \frac{1}{(n-1)!} \int\limits_{x_0}^{x} f(t)(x-t)^{n-1}\,dt. \tag{9.31c}$$

It has to be mentioned here that x_0 is not an additional arbitrary constant, since the change in x_0 results in the change of C_k because of the relation

$$C_k = \frac{1}{(k-1)!} y^{(k-1)}(x_0). \tag{9.31d}$$

9.1.2.3 Linear n-th Order Differential Equations

1. Classification

A differential equation of the form

$$y^{(n)} + a_1 y^{(n-1)} + a_2 y^{(n-2)} + \cdots + a_{n-1} y' + a_n y = F \tag{9.32}$$

is called an n-th order linear differential equation. Here F and the coefficients a_i are functions of x, which are supposed to be continuous in a certain interval. If $a_1, a_2, \dots, a_n$ are constants, it is called a *differential equation with constant coefficients*. If $F \equiv 0$, then the linear differential equation is *homogeneous*, and if $F \not\equiv 0$, then it is *inhomogeneous*.

2. Fundamental System of Solutions

A system of n solutions $y_1, y_2, \dots, y_n$ of a homogeneous linear differential equation is called a *fundamental system* if these functions are *linearly independent* on the considered interval, i.e., their linear combination $C_1 y_1 + C_2 y_2 + \cdots + C_n y_n$ is not identically zero for any system of values $C_1, C_2, \dots, C_n$, except for the values $C_1 = C_2 = \cdots = C_n = 0$. The solutions $y_1, y_2, \dots, y_n$ of a linear homogeneous differential equation form a fundamental system on the considered interval if and only if their *Wronskian determinant*

$$W = \begin{vmatrix} y_1 & y_2 & \cdots & y_n \\ y_1' & y_2' & \cdots & y_n' \\ \dots & \dots & \dots & \dots \\ y_1^{(n-1)} & y_2^{(n-1)} & \cdots & y_n^{(n-1)} \end{vmatrix} \tag{9.33}$$

is non-zero. For every solution system of a homogeneous linear differential equation the *formula of Liouville* is valid:

$$W(x) = W(x_0) \exp\left(-\int_{x_0}^{x} a_{n-1}(x)\,dx\right). \tag{9.34}$$

It follows from (9.34) that if the Wronskian determinant is zero somewhere in the solution interval, then it can be only identically zero. This means: The n solutions $y_1, y_2, \ldots, y_n$ of the homogeneous linear differential equation are linearly dependent if even for a single point x_0 of the considered interval $W(x_0) = 0$. If the solutions $y_1, y_2, \ldots, y_n$ form a fundamental system of the differential equation, then the general solution of the linear homogeneous differential equation (9.32) is given as

$$y = C_1\,y_1 + C_2\,y_2 + \cdots + C_n\,y_n. \tag{9.35}$$

A linear n-th order homogeneous differential equation has exactly n linearly independent solutions on an interval, where the coefficient functions $a_i(x)$ are continuous.

3. Lowering the Order

If a particular solution y_1 of a homogeneous differential equation is known, by assuming

$$y = y_1(x)u(x) \tag{9.36}$$

one can determine further solutions from a homogeneous linear differential equation of order $n-1$ for $u'(x)$.

4. Superposition Principle

If y_1 and y_2 are two solutions of the differential equation (9.32) for different right-hand sides F_1 and F_2, then their sum $y = y_1 + y_2$ is a solution of the same differential equation with the right-hand side $F = F_1 + F_2$. From this observation it follows that to get the general solution of an inhomogeneous differential equation it is sufficient to add any particular solution of the inhomogeneous differential equation to the general solution of the corresponding homogeneous differential equation.

5. Decomposition Theorem

If an inhomogeneous differential equation (9.32) has real coefficients and its right-hand side is complex in the form $F = F_1 + \mathrm{i}F_2$ with some real functions F_1 and F_2, then the solution $y = y_1 + \mathrm{i}y_2$ is also complex, where y_1 and y_2 are the two solutions of the two inhomogeneous differential equations (9.32) with the corresponding right-hand sides F_1 and F_2.

6. Solution of Inhomogeneous Differential Equations (9.32) by Means of Quadratures

If the fundamental system of the corresponding homogeneous differential equation is already known, there are the following two solution methods to continue the calculations:

1. Method of Variation of Constants Looking for the solution in the form

$$y = C_1y_1 + C_2y_2 + \cdots + C_ny_n \tag{9.37a}$$

where $C_1, C_2, \ldots, C_n$, here treated as functions of x. There are infinitely many such functions, but requiring that they satisfy the equations

$$\begin{aligned} C_1'y_1 + C_2'y_2 + \cdots + C_n'y_n &= 0,\\ C_1'y_1' + C_2'y_2' + \cdots + C_n'y_n' &= 0,\\ \cdots\cdots\cdots\cdots\cdots&\\ C_1'y_1^{(n-2)} + C_2'y_2^{(n-2)} + \cdots + C_n'y_n^{(n-2)} &= 0 \end{aligned} \tag{9.37b}$$

and substituting y into (9.32) with these equalities follows

$$C_1'y_1^{(n-1)} + C_2'y_2^{(n-1)} + \cdots + C_n'y_n^{(n-1)} = F. \tag{9.37c}$$

Because the Wronskian determinant of the coefficients in the linear system of equations (9.37b) and (9.37c) is different from zero, one gets a unique solution for the unknown functions $C_1', C_2', \ldots, C_n'$,

and their integrals give the functions $C_1, C_2, \dots, C_n$.

■ $$y'' + \frac{x}{1-x}y' - \frac{1}{1-x}y = x - 1. \tag{9.37d}$$

In the interval $x > 1$ or $x < 1$ all assumptions on the coefficients are fulfilled. First the homogeneous equation $\overline{y}'' + \frac{x}{1-x}\overline{y}' - \frac{1}{1-x}\overline{y} = 0$ is solved. A particular solution is $\varphi_1 = e^x$. Then one looks for a second one in the form $\varphi_2 = e^x u(x)$, and with the notation $u'(x) = v(x)$ one gets the first-order differential equation $v' + \left(1 + \frac{1}{1-x}\right) v = 0$. A solution of this equation is $v(x) = (1-x)e^{-x}$, and therefore, $u(x) = \int v(x)\, dx = \int (1-x)e^{-x}\, dx = xe^{-x}$. With this result $\varphi_2 = x$ is obtained for the second element of the fundamental system. The general solution of the homogeneous equation is $\overline{y}(x) = C_1 e^x + C_2 x$. The variation of constants with $u_1(x)$ and $u_2(x)$ instead of $C_1(x)$ and $C_2(x)$ is now:

$$\begin{aligned} y(x) &= u_1(x)e^x + u_2(x)x, \\ y'(x) &= u_1(x)e^x + u_2(x) + u_1{}'(x)e^x + u_2{}'(x)x, \quad && u_1{}'(x)e^x + u_2{}'(x)x = 0, \\ y''(x) &= u_1(x)e^x + u_1{}'(x)e^x + u_2{}'(x), && u_1{}'(x)e^x + u_2{}'(x) = x - 1, \quad \text{so} \end{aligned}$$

$$u_1{}'(x) = xe^{-x}, \quad u_2{}'(x) = -1, \quad \text{i.e.,} \quad u_1(x) = -(1+x)e^{-x} + C_1, \quad u_2(x) = -x + C_2.$$

With this result the general solution of the inhomogeneous differential equation is:

$$y(x) = -(1+x^2) + C_1 e^x + (C_2 - 1)x = -(1+x^2) + C_1{}^* e^x + C_2{}^* x. \tag{9.37e}$$

2. Method of Cauchy In the general solution

$$y = C_1 y_1 + C_2 y_2 + \cdots + C_n y_n \tag{9.38a}$$

of the homogeneous differential equation associated to (9.32) one determines the constants such that for an arbitrary parameter α the equations $y = 0,\ y' = 0, \dots,\ y^{(n-2)} = 0,\ y^{(n-1)} = F(\alpha)$ are satisfied. In this way one gets a particular solution of the homogeneous equation, denoted by $\varphi(x, \alpha)$, and then

$$y = \int\limits_{x_0}^{x} \varphi(x, \alpha)\, d\alpha \tag{9.38b}$$

is a particular solution of the inhomogeneous differential equation (9.32). This solution and their derivatives up to order $(n-1)$ are equal to zero at the point $x = x_0$.

■ The general solution of the homogeneous equation associated to the differential equation (9.37d) which has been solved by the method variation of constants is $y = C_1 e^x + C_2 x$. From this result follows $y(\alpha) = C_1 e^\alpha + C_2 \alpha = 0$, $y'(\alpha) = C_1 e^\alpha + C_2 = \alpha - 1$ and $\varphi(x, \alpha) = \alpha e^{-\alpha} e^x - x$, so that the particular solution $y(x)$ of the inhomogeneous differential equation with $y(x_0) = y'(x_0) = 0$ is: $y(x) = \int_{x_0}^{x} (\alpha e^{-\alpha} e^x - x)\, d\alpha = (x_0 + 1)e^{x - x_0} + (x_0 - 1)x - x^2 - 1$. With this result one can get the general solution $y(x) = C_1{}^* e^x + C_2{}^* x - (x^2 + 1)$ of the inhomogeneous differential equation.

9.1.2.4 Solution of Linear Differential Equations with Constant Coefficients

1. Operational Notation

The differential equation (9.32) can be written symbolically in the form

$$P_n(D)y \equiv \left(D^n + a_1 D^{n-1} + a_2 D^{n-2} + \cdots + a_{n-1} D + a_n\right) y = F, \tag{9.39a}$$

where D is a differential operator:

$$Dy = \frac{dy}{dx}, \quad D^k y = \frac{d^k y}{dx^k}. \tag{9.39b}$$

If the coefficients a_i are constants, then $P_n(D)$ is a usual polynomial in the operator D of degree n.

2. Solution of the Homogeneous Differential Equation with Constant Coefficients

To determine the general solution of the homogeneous differential equation (9.39a) with $F = 0$, i.e.,

$$P_n(D)y = 0 \tag{9.40a}$$

one has to find the roots $r_1, r_2, \ldots, r_n$ of the characteristic equation

$$P_n(r) = r^n + a_1 r^{n-1} + a_2 r^{n-2} + \cdots + a_{n-1} r + a_n = 0. \tag{9.40b}$$

Every root r_i determines a solution $e^{r_i x}$ of the equation $P_n(D)y = 0$. If a root r_i has a higher multiplicity k, then $xe^{r_i x}$, $x^2 e^{r_i x}, \ldots,$ $x^{k-1} e^{r_i x}$ are also solutions. The linear combination of all these solutions is the general solution of the homogeneous differential equation:

$$y = C_1 e^{r_1 x} + C_2 e^{r_2 x} + \cdots + e^{r_i x}\left(C_i + C_{i+1}x + \cdots + C_{i+k-1}x^{k-1}\right) + \cdots. \tag{9.40c}$$

If the coefficients a_i are all real, then the complex roots of the characteristic equation are pairwise conjugate with the same multiplicity. In this case, for $r_1 = \alpha + \mathrm{i}\beta$ and $r_2 = \alpha - \mathrm{i}\beta$ one can replace the corresponding complex solution functions $e^{r_1 x}$ and $e^{r_2 x}$ by the real functions $e^{\alpha x}\cos\beta x$ and $e^{\alpha x}\sin\beta x$. The resulting expression $C_1\cos\beta x + C_2\sin\beta x$ can be written in the form $A\cos(\beta x + \varphi)$ with some constants A and φ.

■ In the case of the differential equation $y^{(6)} + y^{(4)} - y'' - y = 0$, the characteristic equation is $r^6 + r^4 - r^2 - 1 = 0$ with roots $r_1 = 1$, $r_2 = -1$, $r_{3,4} = \mathrm{i}$, $r_{5,6} = -\mathrm{i}$. The general solution can be given in two forms:

$$y = C_1 e^x + C_2 e^{-x} + (C_3 + C_4 x)\cos x + (C_5 + C_6 x)\sin x, \quad \text{or}$$
$$y = C_1 e^x + C_2 e^{-x} + A_1\cos(x + \varphi_1) + xA_2\cos(x + \varphi_2).$$

3. Hurwitz Theorem

In different applications, e.g., in vibration theory, it is important to know whether a solution of a given homogeneous differential equation with constant coefficients tend to zero for $x \to +\infty$ or not. It tends to zero, obviously, if the real parts of the roots of the characteristic equation (9.40b) are negative. According to the *Hurwitz theorem* an equation

$$a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 = 0 \tag{9.41a}$$

has only roots with negative real part if and only if all the determinants

$$D_1 = a_1, \quad D_2 = \begin{vmatrix} a_1 & a_0 \\ a_3 & a_2 \end{vmatrix}, \quad D_3 = \begin{vmatrix} a_1 & a_0 & 0 \\ a_3 & a_2 & a_1 \\ a_5 & a_4 & a_3 \end{vmatrix}, \ldots, \quad D_n = \begin{vmatrix} a_1 & a_0 & 0 & \ldots & 0 \\ a_3 & a_2 & a_1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & a_n \end{vmatrix}$$

$$(\text{with } a_m = 0 \quad \text{for} \quad m > n) \tag{9.41b}$$

are positive. The determinants D_k have on their diagonal the coefficients $a_1, a_2, \ldots, a_k$ $(k = 1, 2, \ldots, n)$, and the coefficient-indices are decreasing from left to right. Coefficients with negative indices and also with indices larger than n are all put to 0.

■ For a cubic polynomial the determinants have in accordance to (9.41b) the following form:

$$D_1 = a_1, \quad D_2 = \begin{vmatrix} a_1 & a_2 \\ a_3 & a_2 \end{vmatrix}, D_3 = \begin{vmatrix} a_1 & a_0 & 0 \\ a_3 & a_2 & a_1 \\ 0 & 0 & a_3 \end{vmatrix}.$$

4. Solution of Inhomogeneous Differential Equations with Constant Coefficients

These differential equations can be solved by the method variation of constants, or by the method of Cauchy, or with the operator method (see 9.2.2.3, **5.**, p. 588). If the right-hand side of the inhomogeneous differential equation (9.32) has a special form, then a particular solution can be determined easily.

1. Form: $\boldsymbol{F(x) = Ae^{\alpha x}, \quad P_n(\alpha) \neq 0}$ (9.42a)

A particular solution is

$$y = \frac{Ae^{\alpha x}}{P_n(\alpha)}\,. \tag{9.42b}$$

If α is a root of the characteristic equation of multiplicity m, i.e., if

$$P_n(\alpha) = P_n{}'(\alpha) = \ldots = P_n{}^{(m-1)}(\alpha) = 0, \tag{9.42c}$$

then $y = \dfrac{Ax^m e^{\alpha x}}{P_n{}^{(m)}(\alpha)}$ is a particular solution. These formulas can also be used by applying the decomposition theorem, if the right side is

$$F(x) = Ae^{\alpha x}\cos\omega x \quad \text{or} \quad Ae^{\alpha x}\sin\omega x. \tag{9.42d}$$

The corresponding particular solutions are the real or the imaginary part of the solution of the same differential equation for

$$F(x) = Ae^{\alpha x}(\cos\omega x + \mathrm{i}\sin\omega x) = Ae^{(\alpha+\mathrm{i}\omega)x} \tag{9.42e}$$

on the right-hand side.

■ **A:** For the differential equation $y'' - 6y' + 8y = e^{2x}$, the characteristic polynomial is $P(D) = D^2 - 6D + 8$ with $P(2) = 0$ and $P'(D) = 2D - 6$ with $P'(2) = 2\cdot 2 - 6 = -2$, so the particular solution is $y = -\dfrac{xe^{2x}}{2}$.

■ **B:** The differential equation $y'' + y' + y = e^x \sin x$ results in the equation $(D^2 + D + 1)y = e^{(1+\mathrm{i})x}$. From its solution $y = \dfrac{e^{(1+\mathrm{i})x}}{(1+\mathrm{i})^2 + (1+\mathrm{i}) + 1} = \dfrac{e^x(\cos x + \mathrm{i}\sin x)}{2+3\mathrm{i}}$ one gets a particular solution $y_1 = \dfrac{e^x}{13}(2\sin x - 3\cos x)$. Here y_1 is the imaginary part of y.

2. Form: $F(x) = Q_n(x)e^{\alpha x}$, $Q_n(x)$ is a polynomial of degree n (9.43)

A particular solution can always be found in the same form, i.e., as an expression $y = R(x)e^{\alpha x}$. $R(x)$ is a polynomial of degree n multiplied by x^m if α is a root of the characteristic equation with a multiplicity m. Considering the coefficients of the polynomial $R(x)$ as unknowns and substituting the expression into the inhomogeneous differential equation a linear system of equations is obtained for the coefficients, and this system of equations always has a unique solution.

This method is very useful especially in the cases of $F(x) = Q_n(x)$ for $\alpha = 0$ and $F(x) = Q_n(x)e^{rx}\cos\omega x$ or $F(x) = Q_n(x)e^{rx}\sin\omega x$ for $\alpha = r \pm \mathrm{i}\omega$. There is a solution in the form $y = x^m e^{rx}[M_n(x)\cos\omega x + N_n(x)\sin\omega x]$.

■ The roots of the characteristic equation associated to the differential equation $y^{(4)} + 2y''' + y'' = 6x + 2x\sin x$ are $k_1 = k_2 = 0$, $k_3 = k_4 = -1$. Because of the superposition principle (see 9.1.2.3, **4.**, p. 554), one can calculate the particular solutions of the inhomogeneous differential equation for the summands of the right-hand side separately. For the first summand the substitution of the given form $y_1 = x^2(ax+b)$ results in a right-hand side $12a + 2b + 6ax = 6x$, and so: $a = 1$ and $b = -6$. For the second summand one substitutes $y_2 = (cx+d)\sin x + (fx+g)\cos x$. One gets the coefficients by coefficient comparison from $(2g+2f-6c+2fx)\sin x - (2c+2d+6f+2cx)\cos x = 2x\sin x$, so $c = 0,\ d = -3,\ f = 1,\ g = -1$. Therefore, the general solution is $y = c_1 + c_2x - 6x^2 + x^3 + (c_3x + c_4)e^{-x} - 3\sin x + (x-1)\cos x$.

3. Euler Differential Equation

The Euler differential equation

$$\sum_{k=0}^{n} a_k(cx+d)^k y^{(k)} = F(x) \tag{9.44a}$$

can be transformed with the substitution

$$cx + d = e^t \tag{9.44b}$$

into a linear differential equation with constant coefficients.

■ The differential equation $x^2y'' - 5xy' + 8y = x^2$ is a special case of the Euler differential equation for $n = 2$. With the substitution $x = e^t$ it becomes the differential equation discussed earlier in ■ **A**, p. 557: $\frac{d^2y}{dt^2} - 6\frac{dy}{dt} + 8y = e^{2t}$. The general solution is $y = C_1e^{2t} + C_2e^{4t} - \frac{t}{2}e^{2t} = C_1x^2 + C_2x^4 - \frac{x^2}{2}\ln|x|$.

9.1.2.5 Systems of Linear Differential Equations with Constant Coefficients

1. Normal Form

The following simple case of a system of first-order linear differential equations with constant coefficients is called a *normal system* or a *normal form*:

$$\left.\begin{array}{l} {y_1}' = a_{11}y_1 + a_{12}y_2 + \cdots + a_{1n}y_n, \\ {y_2}' = a_{21}y_1 + a_{22}y_2 + \cdots + a_{2n}y_n, \\ \dots\dots\dots\dots\dots\dots\dots\dots \\ {y_n}' = a_{n1}y_1 + a_{n2}y_2 + \cdots + a_{nn}y_n. \end{array}\right\} \tag{9.45a}$$

To find the general solution of such a system, one has to find first the roots of the characteristic equation

$$\begin{vmatrix} a_{11} - r & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} - r & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} - r \end{vmatrix} = 0. \tag{9.45b}$$

To every single root r_i of this equation there is a system of particular solutions

$$y_1 = A_1e^{r_ix}, \quad y_2 = A_2e^{r_ix}, \dots, y_n = A_ne^{r_ix}, \tag{9.45c}$$

whose coefficients A_k $(k = 1, 2, \dots, n)$ are determined from the homogeneous linear equation system

$$\begin{array}{l} (a_{11} - r_i)A_1 + a_{12}A_2 + \cdots + a_{1n}A_n = 0, \\ \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\ a_{n1}A_1 + a_{n2}A_2 + \cdots + (a_{nn} - r_i)A_n = 0. \end{array} \tag{9.45d}$$

This system gives the relations between the values of the coefficients A_k (see Trivial Solution and Fundamental System in 4.5.2.1, **2.**, p. 309). For every r_i, the particular solutions determined this way will contain an arbitrary constant. If all the roots of the characteristic equation are different, the sum of these particular solutions contains n independent arbitrary constants, so in this way one gets the general solution. If a root r_i has a multiplicity m in the characteristic equation, the system of particular solutions corresponding to this root has the form

$$y_1 = A_1(x)e^{r_ix}, \quad y_2 = A_2(x)e^{r_ix}, \dots, y_n = A_n(x)e^{r_ix}, \tag{9.45e}$$

where $A_1(x), \dots, A_n(x)$ are polynomials of degree at most $m - 1$. After substituting these expressions with unknown coefficients of the polynomials $A_k(x)$ into the differential equation system one first can cancel the factor e^{r_ix}, then one compares the coefficients of the different powers of x to have linear equations for the unknown coefficients of the polynomials, and among them m can be chosen freely. In this way, one gets a part of the solution with m arbitrary constants. The degree of the polynomials can be less than $m - 1$.

In the special case when the system (9.45a) is symmetric, i.e., when $a_{ik} = a_{ki}$, then it is sufficient to substitute $A_i(x) = \text{const}$. For complex roots of the characteristic equation, the general solution can be transformed into a real form in the same way as has been shown for the case of a differential equation with constant coefficients (see 9.1.2.4, p. 555).

■ For the system ${y_1}' = 2y_1 + 2y_2 - y_3$, ${y_2}' = -2y_1 + 4y_2 + y_3$, ${y_3}' = -3y_1 + 8y_2 + 2y_3$ the characteristic

equation has the form

$$\begin{vmatrix} 2-r & 2 & -1 \\ -2 & 4-r & 1 \\ -3 & 8 & 2-r \end{vmatrix} = -(r-6)(r-1)^2 = 0.$$

For the simple root $r_1 = 6$ one gets $-4A_1+2A_2-A_3 = 0$, $-2A_1-2A_2+A_3 = 0$, $-3A_1+8A_2-4A_3 = 0$. From this system one has $A_1 = 0$, $A_2 = \frac{1}{2}A_3 = C_1$, $y_1 = 0$, $y_2 = C_1e^{6x}$, $y_3 = 2C_1e^{6x}$. For the multiple root $r_2 = 1$ one gets $y_1 = (P_1x + Q_1)e^x$, $y_2 = (P_2x + Q_2)e^x$, $y_3 = (P_3x + Q_3)e^x$. Substitution into the differential equations yields

$$P_1x + (P_1 + Q_1) = (2P_1 + 2P_2 - P_3)x + (2Q_1 + 2Q_2 - Q_3),$$

$$P_2x + (P_2 + Q_2) = (-2P_1 + 4P_2 + P_3)x + (-2Q_1 + 4Q_2 + Q_3),$$

$$P_3x + (P_3 + Q_3) = (-3P_1 + 8P_2 + 2P_3)x + (-3Q_1 + 8Q_2 + 2Q_3),$$

which implies that $P_1 = 5C_2$, $P_2 = C_2$, $P_3 = 7C_2$, $Q_1 = 5C_3 - 6C_2$, $Q_2 = C_3$, $Q_3 = 7C_3 - 11C_2$. The general solution is $y_1 = (5C_2x + 5C_3 - 6C_2)e^x$, $y_2 = C_1e^{6x} + (C_2x + C_3)e^x$, $y_3 = 2C_1e^{6x} + (7C_2x + 7C_3 - 11C_2)e^x$.

2. Homogeneous Systems of First-Order Linear Differential Equations with Constant Coefficients

have the general form

$$\sum_{k=1}^{n} a_{ik}y_k{}' + \sum_{k=1}^{n} b_{ik}y_k = 0 \quad (i = 1, 2, \ldots, n). \tag{9.46a}$$

If the determinant $\det(a_{ik})$ does not disappear, i.e.,

$$\det(a_{ik}) \neq 0, \tag{9.46b}$$

then the system (9.46a) can be transformed into the normal form (9.45a).
In the case of $\det(a_{ik}) = 0$ further investigations are necessary (see [9.15]).
The solution can be determined from the general form in the same way as shown for the normal form. The characteristic equation has the form

$$\det(a_{ik}r + b_{ik}) = 0. \tag{9.46c}$$

The coefficients A_i in the solution (9.45c) corresponding to a single root r_j are determined from the equation system

$$\sum_{k=1}^{n} (a_{ik}r_j + b_{ik})A_k = 0 \quad (i = 1, 2, \ldots, n). \tag{9.46d}$$

Otherwise the solution method follows the same ideas as in the case of the normal form.

■ The characteristic equation of the two differential equations $5y_1{}'+4y_1-2y_2{}'-y_2 = 0$, $y_1{}'+8y_1-3y_2 = 0$ is:

$$\begin{vmatrix} 5r+4 & -2r-1 \\ r+8 & -3 \end{vmatrix} = 2r^2 + 2r - 4 = 0, \quad r_1 = 1, \quad r_2 = -2.$$

The coefficients A_1 and A_2 for $r_1 = 1$ can be got from the equations $9A_1 - 3A_2 = 0$, $9A_1 - 3A_2 = 0$ so $A_2 = 3A_1 = 3C_1$. For $r_2 = -2$ one gets analogously $\overline{A}_2 = 2\overline{A}_1 = 2C_2$. The general solution is $y_1 = C_1e^x + C_2e^{-2x}$, $y_2 = 3C_1e^x + 2C_2e^{-2x}$.

3. Inhomogeneous Systems of First-Order Linear Differential Equations

have the general form

$$\sum_{k=1}^{n} a_{ik}y_k{}' + \sum_{k=1}^{n} b_{ik}y_k = F_i(x) \quad (i = 1, 2, \ldots, n). \tag{9.47}$$

1. Superposition Principle If ${y_j}^{(1)}$ and ${y_j}^{(2)}$ $(j = 1, 2, \ldots, n)$ are solutions of inhomogeneous systems which differ from each other only in their right-hand sides ${F_i}^{(1)}$ and ${F_i}^{(2)}$, then the sum $y_j = {y_j}^{(1)} + {y_j}^{(2)}$ $(j = 1, 2, \ldots, n)$ is a solution of this system with the right-hand side $F_i(x) = {F_i}^{(1)}(x) + {F_i}^{(2)}(x)$. Because of this, to get the general solution of an inhomogeneous system it is enough to add a particular solution to the general solution of the corresponding homogeneous system.

2. The Variation of Constants can be used to get a particular solution of the inhomogeneous differential equation system. To do this one uses the general solution of the homogeneous system, and considers the constants $C_1, C_2, \ldots, C_n$ as unknown functions $C_1(x), C_2(x), \ldots, C_n(x)$. Then it is to be substituted into the inhomogeneous system. In the expressions of the derivatives of ${y_k}'$ there is the derivative of the new unknown functions $C_k(x)$. Because $y_1, y_2, \ldots, y_n$ are solutions of the homogeneous system, the terms containing the new unknown functions will be canceled; only their derivatives remain in the equations. This gives for the functions ${C_k}'(x)$ an inhomogeneous linear algebraic equation system which always has a unique solution. After n integrations one gets the functions $C_1(x), C_2(x), \ldots, C_n(x)$. Substituting them into the solution of the homogeneous system instead of the constants results in the particular solution of the inhomogeneous system.

■ For the system of two inhomogeneous differential equations $5{y_1}' + 4y_1 - 2{y_2}' - y_2 = e^{-x}$, ${y_1}' + 8y_1 - 3y_2 = 5e^{-x}$ the general solution of the homogeneous system is (see p. 559) $y_1 = C_1e^x + C_2e^{-2x}$, $y_2 = 3C_1e^x + 2C_2e^{-2x}$. Considering the constants C_1 and C_2 as functions of x and substituting into the original equations gives $5{C_1}'e^x + 5{C_2}'e^{-2x} - 6{C_1}'e^x - 4{C_2}'e^{-2x} = e^{-x}$, ${C_1}'e^x + {C_2}'e^{-2x} = 5e^{-x}$ or ${C_2}'e^{-2x} - {C_1}'e^{-x} = e^{-x}$, ${C_1}'e^x + {C_2}'e^{-2x} = 5e^{-x}$. Therefore, $2{C_1}'e^x = 4e^{-x}$, $C_1 = -e^{-2x} + \text{const}$, $2{C_2}'e^{-2x} = 6e^{-x}$, $C_2 = 3e^x + \text{const}$. Since a particular solution is searched for, one can replace every constant by zero and the result is $y_1 = 2e^{-x}$, $y_2 = 3e^{-x}$. The general solution is finally $y_1 = 2e^{-x} + C_1e^x + C_2e^{-2x}$, $y_2 = 3e^{-x} + 3C_1e^x + 2C_2e^{-2x}$.

3. The Method of Unknown Coefficients is especially useful if on the right-hand side there are special functions in the form $Q_n(x)e^{\alpha x}$. The application is similar to the one, used for differential equations of n-th order (see 9.1.2.5, p. 558).

4. Second-Order Systems

The methods introduced above can also be used for differential equations of higher order. For the system

$$\sum_{k=1}^{n} a_{ik}{y_k}'' + \sum_{k=1}^{n} b_{ik}{y_k}' + \sum_{k=1}^{n} c_{ik}y_k = 0 \quad (i = 1, 2, \ldots, n) \tag{9.48}$$

one can determine particular solutions in the form $y_i = A_ie^{r_ix}$. To do this, one gets r_i from the characteristic equation $\det(a_{ik}r^2 + b_{ik}r + c_{ik}) = 0$, and A_i from the corresponding linear homogeneous algebraic equations.

9.1.2.6 Linear Second-Order Differential Equations

Many special differential equations belong to this class, which often occur in practical applications. Several of them are discussed in this paragraph. For more details of representation, properties and solution methods see [9.15].

1. General Methods

1. Solving the Inhomogeneous Differential Equation by the Help of the Superposition Principle

$$y'' + p(x)y' + q(x)y = F(x). \tag{9.49a}$$

To get the general solution of an inhomogeneous differential equation it is enough to add a particular solution of the inhomogeneous equation to the general solution of the corresponding homogeneous equation.

a) The general solution of the corresponding homogeneous differential equation, i.e., with $F(x) \equiv 0$, is

$$y = C_1y_1 + C_2y_2. \tag{9.49b}$$

Here y_1 and y_2 are two linearly independent particular solutions of (9.49a) (see 9.1.2.3, **2.**, p. 553). If a particular solution y_1 is already known, then the second one y_2 can be determined by the equation (9.34) of Liouville. From (9.34) follows:

$$\begin{vmatrix} y_1 & y_2 \\ y_1{}' & y_2{}' \end{vmatrix} = y_1 y_2{}' - y_1{}' y_2 = y_1^2 \frac{y_1 y_2{}' - y_1{}' y_2}{y_1^2} = y_1^2 \left(\frac{y_2}{y_1}\right)' = A \exp\left(-\int p(x)\, dx\right) \tag{9.49c}$$

giving

$$y_2 = A y_1 \int \frac{\exp\left(-\int p\, dx\right)}{y_1{}^2}\, dx \tag{9.49d}$$

where A can be chosen arbitrarily.

b) A particular solution of the inhomogeneous equation can be determined by the formula

$$y = \frac{1}{A} \int\limits_{x_0}^{x} F(\xi) \exp\left(\int p(\xi)\, d\xi\right) [y_2(x) y_1(\xi) - y_1(x) y_2(\xi)]\, d\xi, \tag{9.49e}$$

where y_1 and y_2 are two particular solutions of the corresponding homogeneous differential equation.

c) A particular solution of the inhomogeneous differential equation can be determined also by variation of constants (see 9.1.2.3, **6.**, p. 554).

2. Solving the Inhomogeneous Differential Equation by the Method of Undetermined Coefficients

$$s(x)y'' + p(x)y' + q(x)y = F(x) \tag{9.50a}$$

If the functions $s(x)$, $p(x)$, $q(x)$ and $F(x)$ are polynomials or functions which can be expanded into a convergent power series around x_0 in a certain domain, where $s(x_0) \neq 0$, then the solutions of this differential equation can also be expanded into a similar series, and these series are convergent in the same domain. Here they should be determined by the method of undetermined coefficients: The solution to be looking for as a series has the form

$$y = a_0 + a_1(x - x_0) + a_2(x - x_0)^2 + \cdots, \tag{9.50b}$$

and it has to be substituted into the differential equation (9.50a). Equating corresponding coefficients (of the same powers of $(x - x_0)$) results in equations to determine the coefficients $a_0, a_1, a_2, \ldots$.

■ To solve the differential equation $y'' + xy = 0$ one substitutes $y = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \cdots$, $y' = a_1 + 2a_2 x + 3a_3 x^2 + \cdots$, and $y'' = 2a_2 + 6a_3 x + \cdots$ getting $2a_2 = 0$, $6a_3 + a_0 = 0, \ldots$. The solution of these equations is $a_2 = 0$, $a_3 = -\dfrac{a_0}{2 \cdot 3}$, $a_4 = -\dfrac{a_1}{3 \cdot 4}$, $a_5 = 0, \ldots$, so the solution is

$$y = a_0 \left(1 - \frac{x^3}{2 \cdot 3} + \frac{x^6}{2 \cdot 3 \cdot 5 \cdot 6} - \cdots\right) + a_1 \left(x - \frac{x^4}{3 \cdot 4} + \frac{x^7}{3 \cdot 4 \cdot 6 \cdot 7} - \cdots\right).$$

3. The Homogeneous Differential Equation

$$x^2 y'' + x p(x) y' + q(x) y = 0 \tag{9.51a}$$

can be solved by the method of undetermined coefficients if the functions $p(x)$ and $q(x)$ can be expanded as a convergent power series of x. The solutions have the form

$$y = x^r (a_0 + a_1 x + a_2 x^2 + \cdots), \tag{9.51b}$$

whose exponent r can be determined from the *defining equation*

$$r(r - 1) + p(0) r + q(0) = 0. \tag{9.51c}$$

If the roots of this equation are different and their difference is not an integer number, then one gets two linearly independent solutions of (9.51a). Otherwise the method of undetermined coefficients results only one solution. Then with the help of (9.49b) one can get a second solution or at least one can find a form which gives a second solution with the method of undetermined coefficients.

■ For the Bessel differential equation (9.52a) one gets only one solution with the method of the un-

determined coefficients in the form $y_1 = \sum_{k=0}^{\infty} a_k x^{n+2k}$ $(a_0 \neq 0)$, which coincides with $J_n(x)$ up to a constant factor. Since $\exp\left(-\int p\,dx\right) = \frac{1}{x}$ one finds a second solution by using formula (9.49d)

$$y_2 = Ay_1 \int \frac{dx}{x \cdot x^{2n}\left(\sum a_k x^{2k}\right)^2} = Ay_1 \int \frac{\sum_{k=0}^{\infty} c_k x^{2k}}{x^{2n+1}} dx = By_1 \ln x + x^{-n} \sum_{k=0}^{\infty} d_k x^{2k}.$$

The determination of the unknown coefficients c_k and d_k is difficult from the a_k's. But this last expression can be used to get the solution with the method of undetermined coefficients. Obviously this form is a series expansion of the function $Y_n(x)$ (9.53c).

2. Bessel Differential Equation

$$x^2y'' + xy' + (x^2 - n^2)y = 0. \tag{9.52a}$$

1. The Defining Equation is in this case

$$r(r-1) + r - n^2 \equiv r^2 - n^2 = 0, \tag{9.52b}$$

so, $r = \pm n$. Substituting

$$y = x^n(a_0 + a_1 x + \cdots) \tag{9.52c}$$

into this equation and equating the coefficients of x^{n+k} to zero gives

$$k(2n+k)a_k + a_{k-2} = 0. \tag{9.52d}$$

For $k = 1$ follows $(2n+1)a_1 = 0$. For the values $k = 2, 3, \ldots$ one obtains

$$a_{2m+1} = 0 \quad (m = 1, 2, \ldots), \quad a_2 = -\frac{a_0}{2(2n+2)},$$

$$a_4 = \frac{a_0}{2 \cdot 4 \cdot (2n+2)(2n+4)}, \ldots, \quad a_0 \text{ is arbitrary.} \tag{9.52e}$$

2. Bessel or Cylindrical Functions The series obtained above for $a_0 = \frac{1}{2^n \Gamma(n+1)}$, where Γ is the gamma function (see 8.2.5, **6.**, p. 514), is a particular solution of the Bessel differential equation (9.52a) for integer values of n. It defines the *Bessel* or *cylindrical function* of the first kind of index n

$$J_n(x) = \frac{x^n}{2^n \Gamma(n+1)} \left(1 - \frac{x^2}{2(2n+2)} + \frac{x^4}{2 \cdot 4 \cdot (2n+2)(2n+4)} - \cdots\right)$$

$$= \sum_{k=0}^{\infty} \frac{(-1)^k \left(\frac{x}{2}\right)^{n+2k}}{k!\,\Gamma(n+k+1)}. \tag{9.53a}$$

The graphs of functions J_0 and J_1 are shown in **Fig. 9.12**.
The general solution of the Bessel differential equation for non-integer n has the form

$$y = C_1 J_n(x) + C_2 J_{-n}(x), \tag{9.53b}$$

where $J_{-n}(x)$ is defined by the infinite series obtained from the series representation of $J_n(x)$ by replacing n with $-n$. For integer n, holds $J_{-n}(x) = (-1)^n J_n(x)$. In this case, the term $J_{-n}(x)$ in the general solution should be replaced with the Bessel function of the second kind

$$Y_n(x) = \lim_{m \to n} \frac{J_m(x) \cos m\pi - J_{-m}(x)}{\sin m\pi}, \tag{9.53c}$$

which is also called the *Weber function*. For the series expansion of $Y_n(x)$ see, e.g., [9.15]. The graphs of the functions Y_0 and Y_1 are shown in **Fig. 9.13**.

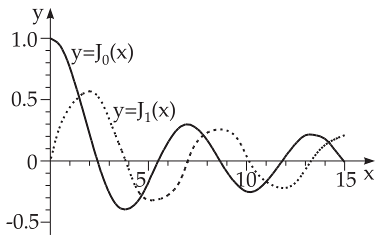


Figure 9.12

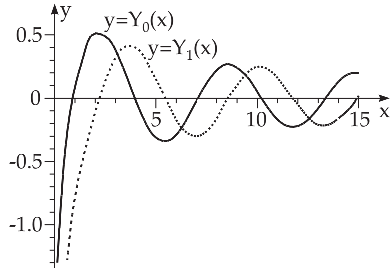


Figure 9.13

3. Bessel Functions with Imaginary Variables In some applications one uses Bessel functions with pure imaginary variables. In this case it is to be considered the product $\mathrm{i}^{-n}J_n(\mathrm{i}x)$ which will be denoted by $I_n(x)$:

$$I_n(x) = \mathrm{i}^{-n}J_n(\mathrm{i}x) = \frac{\left(\frac{x}{2}\right)^n}{\Gamma(n+1)} + \frac{\left(\frac{x}{2}\right)^{n+2}}{1!\Gamma(n+2)} + \frac{\left(\frac{x}{2}\right)^{n+4}}{2!\Gamma(n+3)} + \cdots. \tag{9.54a}$$

The functions $I_n(x)$ are solutions of the differential equation

$$x^2y'' + xy' - (x^2 + n^2)y = 0. \tag{9.54b}$$

A second solution of this differential equation is the *MacDonald function*

$$K_n(x) = \frac{\pi}{2}\frac{I_{-n}(x) - I_n(x)}{\sin n\pi}. \tag{9.54c}$$

If n converges to an integer number, this expression also converges.

The functions $I_n(x)$ and $K_n(x)$ are called *modified Bessel functions*. The graphs of functions I_0 and I_1 are shown in **Fig. 9.14**; the graphs of functions K_0 and K_1 are illustrated in **Fig. 9.15**. The values of functions $J_0(x)$, $J_1(x)$, $Y_0(x)$, $Y_1(x)$, $I_0(x)$, $I_1(x)$, $K_0(x)$, $K_1(x)$ are given in **Table 21.11**, p. 1106.

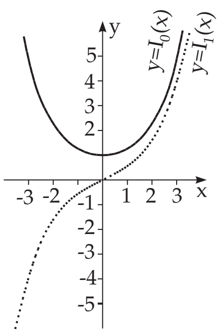


Figure 9.14

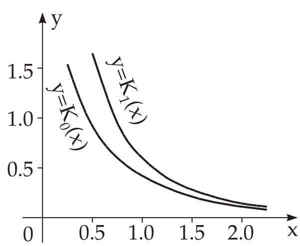


Figure 9.15

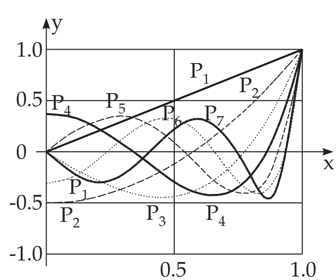


Figure 9.16

4. Important Formulas for the Bessel Functions $J_n(x)$

$$J_{n-1}(x) + J_{n+1}(x) = \frac{2n}{x}J_n(x), \quad \frac{dJ_n(x)}{dx} = -\frac{n}{x}J_n(x) + J_{n-1}(x). \tag{9.55a}$$

The formulas (9.55a) are also valid for the Weber functions $Y_n(x)$.

$$I_{n-1}(x) - I_{n+1}(x) = \frac{2nI_n(x)}{x}, \quad \frac{dI_n(x)}{dx} = I_{n-1}(x) - \frac{n}{x}I_n(x), \tag{9.55b}$$

$$K_{n+1}(x) - K_{n-1}(x) = \frac{2nK_n(x)}{x}, \quad \frac{dK_n(x)}{dx} = -K_{n-1}(x) - \frac{n}{x}K_n(x). \tag{9.55c}$$

For integer numbers n the following formulas are valid:

$$J_{2n}(x) = \frac{2}{\pi}\int_0^{\pi/2} \cos(x\sin\varphi)\cos 2n\varphi\, d\varphi, \tag{9.55d}$$

$$J_{2n+1}(x) = \frac{2}{\pi}\int_0^{\pi/2} \sin(x\sin\varphi)\sin(2n+1)\varphi\, d\varphi \tag{9.55e}$$

or, in complex form,

$$J_n(x) = \frac{-(\mathrm{i})^n}{\pi}\int_0^{\pi} e^{\mathrm{i}x\cos\varphi}\cos n\varphi\, d\varphi. \tag{9.55f}$$

The functions $J_{n+1/2}(x)$ can be expressed by using elementary functions. In particular,

$$J_{1/2}(x) = \sqrt{\frac{2}{\pi x}}\sin x, \tag{9.56a}$$ $$J_{-1/2}(x) = \sqrt{\frac{2}{\pi x}}\cos x. \tag{9.56b}$$

By applying the recursion formulas (9.55a)–(9.55f) the expression for $J_{n+1/2}(x)$ for arbitrary integer n can be given. For large values of x the following asymptotic formulas are valid:

$$J_n(x) = \sqrt{\frac{2}{\pi x}}\left[\cos\left(x - \frac{n\pi}{2} - \frac{\pi}{4}\right) + O\left(\frac{1}{x}\right)\right], \tag{9.57a}$$

$$I_n(x) = \frac{e^x}{\sqrt{2\pi x}}\left[1 + O\left(\frac{1}{x}\right)\right], \tag{9.57b}$$

$$Y_n(x) = \sqrt{\frac{2}{\pi x}}\left[\sin\left(x - \frac{n\pi}{2} - \frac{\pi}{4}\right) + O\left(\frac{1}{x}\right)\right], \tag{9.57c}$$

$$K_n(x) = \sqrt{\frac{\pi}{2x}}e^{-x}\left[1 + O\left(\frac{1}{x}\right)\right]. \tag{9.57d}$$

The expression $O\left(\frac{1}{x}\right)$ means an infinitesimal quantity of the same order as $\frac{1}{x}$ (see the Landau symbol, 2.1.4.9, p. 57).
For further properties of the Bessel functions see [21.1].

5. Important Formulas for the Spherical Bessel Functions Spherical Bessel functions of the first and second kind $j_l(z)$ and $n_l(z)$ follow from the Bessel functions of the first and second kind $J_n(z)$ (9.53a)and $Y_n(z)$ (9.53c) for half odd order index $n = \frac{1}{2}, \frac{3}{2}, \ldots$ as in $j_l(z) = \sqrt{\frac{\pi}{2z}}J_{l+\frac{1}{2}}(z)$ and $n_l(z) = \sqrt{\frac{\pi}{2z}}Y_{l+\frac{1}{2}}(z)$ with $l = 0, 1, 2, \ldots$. They occur as regular or singular solutions of the potential

free radial Schroedinger equation (see 9.2.4.6,**3.**, (9.137b), p. 599) with $V(r) = 0$, $E = \frac{\hbar^2 k^2}{2m}$, $z = kr$ and $s_l(z) = R_l(r)$:

$$z\frac{d^2}{dz^2}[z\, s_l(z)] + [z^2 - l(l+1)]s_l(z) = 0\,, \quad s_l(z) = j_l(z) \text{ or } n_l(z). \tag{9.58a}$$

They also occur in the quantum mechanical scattering theory, where the $n_l(z)$ are called the spherical von Neumann functions. By the help of the Rayleigh formulas

$$j_l(z) = (-z)^l \left(\frac{d}{zdz}\right)^l \frac{\sin z}{z}, \quad n_l(z) = (-z)^l \left(\frac{d}{zdz}\right)^l (-1)\frac{\cos z}{z} \tag{9.58b}$$

follows, so

$$j_0(z) = \frac{\sin z}{z}, \quad j_1(z) = \frac{\sin z - z\cos z}{z^2}, \cdots, \tag{9.58c}$$

$$n_0(z) = -\frac{\cos z}{z}, \quad n_1(z) = -\frac{\cos z + z\sin z}{z^2}, \cdots. \tag{9.58d}$$

Complex spherical functions are used with $\Phi_m(\varphi) = \mathrm{e}^{\mathrm{i}m\varphi}$ in the form $Y_L(\vec{\mathbf{e}}) = \Theta_l^m(\vartheta)\Phi_m(\varphi)$, e.g. in 9.2.4.6, (9.136e), p. 599. In the combined index $L = (l, m)$ $l = 0, 1, 2, \ldots$ denotes the quantum number of the orbital angular momentum. The magnetic quantum number m is restricted to the $2l + 1$ values $m = -l, -l + 1, \cdots, +l$. With the abbreviations

$$j_L(k\vec{\mathbf{r}}) = j_l(kr)Y_L(\vec{\mathbf{e}}_r)\,, \quad n_L(k\vec{\mathbf{r}}) = n_l(kr)Y_L(\vec{\mathbf{e}}_r), \quad \vec{\mathbf{e}}_r = \frac{\vec{\mathbf{r}}}{r} \tag{9.59a}$$

one gets the Kasterinian formulas

$$\mathrm{i}^l j_L(k\vec{\mathbf{r}}) = Y_L\left(\frac{\nabla}{\mathrm{i}k}\right)\frac{\sin kr}{kr}, \quad \mathrm{i}^l n_L(k\vec{\mathbf{r}}) = Y_L\left(\frac{\nabla}{\mathrm{i}k}\right)(-1)\frac{\cos kr}{kr}\,, \tag{9.59b}$$

where ∇ denotes the nabla operator (s. 13.2.6.1, S. 715). The expansion of a plane wave in terms of spherical or Bessel functions gives

$$\mathrm{e}^{\mathrm{i}\vec{\mathbf{k}}\vec{\mathbf{r}}} = 4\pi \sum_L \mathrm{i}^l j_L(k\vec{\mathbf{r}})Y_L^*(\vec{\mathbf{e}}_k), \quad \vec{\mathbf{e}}_k = \frac{\vec{\mathbf{k}}}{k}, \quad \sum_L \cdots = \sum_{l=0}^{\infty}\sum_{m=-l}^{+l} \cdots. \tag{9.59c}$$

There are the following addition theorems

$$\mathrm{i}^l j_L(k(\vec{\mathbf{r}}_1 + \vec{\mathbf{r}}_2)) = 4\pi \sum_{L_1, L_2} C_{LL_1L_2}\mathrm{i}^{l_1+l_2} j_{L_1}(k\vec{\mathbf{r}}_1)j_{L_2}^*(k\vec{\mathbf{r}}_2)\,, \quad r_{1,2} = \text{arbitrarily}, \tag{9.59d}$$

$$\mathrm{i}^l n_L(k(\vec{\mathbf{r}}_1 + \vec{\mathbf{r}}_2)) = 4\pi \sum_{L_1, L_2} C_{LL_1L_2}\mathrm{i}^{l_1+l_2} n_{L_1}(k\vec{\mathbf{r}}_1)j_{L_2}^*(k\vec{\mathbf{r}}_2)\,, \quad r_1 > r_2 \tag{9.59e}$$

with the Clebsch-Gordan coefficients (see 5.3.4.7, p. 345)

$$C_{LL_1L_2} = \int d^2 e\, Y_L(\vec{\mathbf{e}})Y_{L_1}^*(\vec{\mathbf{e}})Y_{L_2}^*(\vec{\mathbf{e}})\,. \tag{9.59f}$$

For further details see [21.1], [9.28] until [9.31].

3. Legendre Differential Equation

Restricting the investigations in this book to the case of real variables and integer parameters $n = 0, 1, 2, \ldots$ the Legendre differential equation has the form

$$(1 - x^2)y'' - 2xy' + n(n+1)y = 0 \quad \text{or} \quad ((1 - x^2)y')' + n(n+1)y = 0. \tag{9.60a}$$

1. Legendre Polynomials or Spherical Harmonics of the First Kind are the particular solutions of the Legendre differential equation for integer n, which can be expanded into the power series $y = \sum_{\nu=0}^{\infty} a_\nu x^\nu$. The method of undetermined coefficients yields the polynomials

$$P_n(x) = \frac{(2n)!}{2^n (n!)^2} \left[x^n - \frac{n(n-1)}{2(2n-1)} x^{n-2} + \frac{n(n-1)(n-2)(n-3)}{2 \cdot 4(2n-1)(2n-3)} x^{n-4} - + \cdots \right],$$

$$(|x| < \infty; n = 0, 1, 2, \ldots). \quad (9.60b)$$

$$P_n(x) = F\left(n+1, -n, 1; \frac{1-x}{2}\right) = \frac{1}{2^n n!} \frac{d^n (x^2-1)^n}{dx^n}, \quad (9.60c)$$

where F denotes the hypergeometric series (see **4.**, p. 567). The first eight polynomials have the following simple form (see 21.12, p. 1108):

$$P_0(x) = 1, \quad (9.60d) \qquad P_1(x) = x, \quad (9.60e)$$

$$P_2(x) = \frac{1}{2}(3x^2 - 1), \quad (9.60f) \qquad P_3(x) = \frac{1}{2}(5x^3 - 3x), \quad (9.60g)$$

$$P_4(x) = \frac{1}{8}(35x^4 - 30x^2 + 3), \quad (9.60h) \qquad P_5(x) = \frac{1}{8}(63x^5 - 70x^3 + 15x), \quad (9.60i)$$

$$P_6(x) = \frac{1}{16}(231x^6 - 315x^4 + 105x^2 - 5), \quad (9.60j) \qquad P_7(x) = \frac{1}{16}(429x^7 - 693x^5 + 315x^3 - 35x). \quad (9.60k)$$

The graphs of $P_n(x)$ for the values from $n = 1$ to $n = 7$ are represented in **Fig. 9.16**. The numerical values can be calculated easily by pocket calculators or from function tables.

2. Properties of the Legendre Polynomials of the First Kind

a) Integral Representation:

$$P_n(x) = \frac{1}{\pi} \int_0^\pi (x \pm \cos\varphi \sqrt{x^2-1})^n \, d\varphi = \frac{1}{\pi} \int_0^\pi \frac{d\varphi}{(x \pm \cos\varphi \sqrt{x^2-1})^{n+1}}. \quad (9.61a)$$

The signs can be chosen arbitrarily in both equations.

b) Recursion Formulas:

$$(n+1)P_{n+1}(x) = (2n+1)xP_n(x) - nP_{n-1}(x) \quad (n \geq 1; P_0(x) = 1,\ P_1(x) = x), \quad (9.61b)$$

$$(x^2 - 1)\frac{dP_n(x)}{dx} = n[xP_n(x) - P_{n-1}(x)] \quad (n \geq 1). \quad (9.61c)$$

c) Orthogonality Relation:

$$\int_{-1}^{1} P_n(x) P_m(x) \, dx = \begin{cases} 0 & \text{for } m \neq n, \\ \dfrac{2}{2n+1} & \text{for } m = n. \end{cases} \quad (9.61d)$$

d) Root Theorem: All the n roots of $P_n(x)$ are real and single and are in the interval $(-1, 1)$.

e) Generating Function: The Legendre polynomial of the first kind can be represented as the power series expansion of the function

$$\frac{1}{\sqrt{1 - 2rx + r^2}} = \sum_{n=0}^{\infty} P_n(x) r^n. \quad (9.61e)$$

For further properties of the Legendre polynomials of the first kind see [21.1].

3. Legendre Functions or Spherical Harmonics of the Second Kind A second particular solution $Q_n(x)$ can be got, which is valid for $|x| > 1$ and linearly independent of $P_n(x)$, see (9.61a), by the power series expansion $\sum_{\nu=-\infty}^{-(n+1)} b_\nu x^\nu$:

$$Q_n(x) = \frac{2^n(n!)^2}{(2n+1)!}x^{-(n+1)}F\left(\frac{n+1}{2}, \frac{n+2}{2}, \frac{2n+3}{2}; \frac{1}{x^2}\right)$$
$$= \frac{2^n(n!)^2}{(2n+1)!}\left[x^{-(n+1)} + \frac{(n+1)(n+2)}{2(2n+3)}x^{-(n+3)}\right.$$
$$\left. + \frac{(n+1)(n+2)(n+3)(n+4)}{2\cdot 4\cdot(2n+3)(2n+5)}x^{-(n+5)} + \cdots\right]. \tag{9.62a}$$

The representation of $Q_n(x)$ valid for $|x| < 1$ is:

$$Q_n(x) = \frac{1}{2}P_n(x)\ln\frac{1+x}{1-x} - \sum_{k=1}^{n}\frac{1}{k}P_{k-1}(x)P_{n-k}(x). \tag{9.62b}$$

The spherical harmonics of the first and second kind are also called the *associated Legendre functions* (see also 9.2.4.6, **4.**, (9.138c), p. 600).

4. Hypergeometric Differential Equation

The *hypergeometric differential equation* is the equation

$$x(1-x)\frac{d^2y}{dx^2} + [\gamma - (\alpha+\beta+1)x]\frac{dy}{dx} - \alpha\beta y = 0, \tag{9.63a}$$

where α, β, γ are parameters. It contains several important special cases.

a) For $\alpha = n+1$, $\beta = -n$, $\gamma = 1$, and $x = \frac{1-z}{2}$ it is the Legendre differential equation.

b) If $\gamma \neq 0$ or γ is not a negative integer, it has the hypergeometric series or hypergeometric function as a particular solution :

$$F(\alpha,\beta,\gamma;x) = 1 + \frac{\alpha\cdot\beta}{1\cdot\gamma}x + \frac{\alpha(\alpha+1)\beta(\beta+1)}{1\cdot 2\cdot\gamma(\gamma+1)}x^2 + \cdots$$
$$+ \frac{\alpha(\alpha+1)\dots(\alpha+n)\beta(\beta+1)\dots(\beta+n)}{1\cdot 2\dots(n+1)\cdot\gamma(\gamma+1)\dots(\gamma+n)}x^{n+1} + \cdots, \tag{9.63b}$$

which is absolutely convergent for $|x| < 1$. The convergence for $x = \pm 1$ depends on the value of $\delta = \gamma - \alpha - \beta$. For $x = 1$ it is convergent if $\delta > 0$, it is divergent if $\delta \leq 0$. For $x = -1$ it is absolutely convergent if $\delta < 0$, it is conditionally convergent for $-1 < \delta \leq 0$, and it is divergent for $\delta \leq -1$.

c) For $2-\gamma \neq 0$ or not equal to a negative integer it has a particular solution

$$y = x^{1-\gamma}F(\alpha+1-\gamma, \beta+1-\gamma, 2-\gamma, x). \tag{9.63c}$$

d) In some special cases the hypergeometric series can be reduced to elementary functions, e.g.,

$$F(1,\beta,\beta;x) = F(\alpha,1,\alpha;x) = \frac{1}{1-x}, \tag{9.64a}$$
$$F(-n,\beta,\beta;-x) = (1+x)^n, \tag{9.64b}$$

$$F(1,1,2;-x) = \frac{\ln(1+x)}{x}, \tag{9.64c}$$
$$F\left(\frac{1}{2},\frac{1}{2},\frac{3}{2};x^2\right) = \frac{\arcsin x}{x}, \tag{9.64d}$$

$$\lim_{\beta\to\infty} F\left(1,\beta,1;\frac{x}{\beta}\right) = e^x. \tag{9.64e}$$

5. Laguerre Differential Equation

Restricting the investigation to integer parameters ($n = 0, 1, 2, \ldots$) and real variables, the *Laguerre differential equation* has the form

$$xy'' + (\alpha + 1 - x)y' + ny = 0. \tag{9.65a}$$

Particular solutions are the *Laguerre polynomials*

$$L_n^{(\alpha)}(x) = \frac{e^x x^{-\alpha}}{n!}\frac{d^n}{dx^n}(e^{-x}x^{n+\alpha}) = \sum_{k=0}^{n}\binom{n+\alpha}{n-k}\frac{(-x)^k}{k!}. \tag{9.65b}$$

The recursion formula for $n \geq 1$ is:

$$(n+1)L_{n+1}^{(\alpha)}(x) = (-x + 2n + \alpha + 1)L_n^{(\alpha)}(x) - (n+\alpha)L_{n-1}^{(\alpha)}(x), \tag{9.65c}$$

$$L_0^{(\alpha)}(x) = 1, \quad L_1^{(\alpha)} = 1 + \alpha - x. \tag{9.65d}$$

An orthogonality relation for $\alpha > -1$ holds:

$$\int_0^\infty e^{-x}x^\alpha L_m^{(\alpha)}(x)L_n^{(\alpha)}(x)\,dx = \begin{cases} 0 & \text{for } m \neq n, \\ \binom{n+\alpha}{n}\Gamma(1+\alpha) & \text{for } m = n. \end{cases} \tag{9.65e}$$

Γ denotes the gamma function (see 8.2.5, **6.**, p. 514).

6. Hermite Differential Equation

Two defining equations are often used in the literature:

a) Defining Equation of Type 1:

$$y'' - xy' + ny = 0 \quad (n = 0, 1, 2, \ldots). \tag{9.66a}$$

b) Defining Equation of Type 2:

$$y'' - 2xy' + ny = 0 \quad (n = 0, 1, 2, \ldots). \tag{9.66b}$$

Particular solutions are the *Hermite polynomials*, $He_n(x)$ for the defining equation of type 1, and $H_n(x)$ for the defining equation of type 2.

a) Hermite Polynomials for Defining Equation of Type 1:

$$\begin{aligned} He_n(x) &= (-1)^n \exp\left(\frac{x^2}{2}\right)\frac{d^n}{dx^n}\exp\left(-\frac{x^2}{2}\right) \\ &= x^n - \binom{n}{2}x^{n-2} + 1\cdot 3\binom{n}{4}x^{n-4} - 1\cdot 3\cdot 5\binom{n}{6}x^{n-6} + \cdots \quad (n \in \mathbb{N}). \end{aligned} \tag{9.66c}$$

For $n \geq 1$ the following recursion formulas are valid:

$$He_{n+1}(x) = xHe_n(x) - nHe_{n-1}(x), \tag{9.66d}$$ $$He_0(x) = 1, \quad He_1(x) = x. \tag{9.66e}$$

The orthogonality relation is:

$$\int_{-\infty}^{+\infty} \exp\left(-\frac{x^2}{2}\right) He_m(x)He_n(x)\,dx = \begin{cases} 0 & \text{for } m \neq n, \\ n!\sqrt{2\pi} & \text{for } m = n. \end{cases} \tag{9.66f}$$

b) Hermite Polynomials for Defining Equation of Type 2:

$$H_n(x) = (-1)^n \exp\left(x^2\right)\frac{d^n}{dx^n}\exp\left(-x^2\right) \quad (n \in \mathbb{N}). \tag{9.66g}$$

The relation with the Hermite polynomials for defining equation of type 1 is the following:

$$He_n(x) = 2^{-n/2}H_n\left(\frac{x}{\sqrt{2}}\right) \quad (n \in \mathbb{N}). \tag{9.66h}$$

9.1.3 Boundary Value Problems

9.1.3.1 Problem Formulation

1. Notion of the Boundary Value Problem

In different applications, e.g., in mathematical physics, differential equations must be solved as so-called *boundary value problems* (see 9.2.3, p. 589), where the required solution must satisfy previously given relations at the endpoints of an interval of the independent variable. A special case is the linear boundary value problem, where a solution of a linear differential equation should satisfy linear boundary value conditions. In the following section the discussion is restricted to second-order linear differential equations with linear boundary values.

2. Self-Adjoint Differential Equation

Self-adjoint differential equations are important special second-order differential equations of the form

$$[py']' - qy + \lambda \varrho y = f. \tag{9.67a}$$

The linear boundary values are the homogeneous conditions

$$A_0 y(a) + B_0 y'(a) = 0, \quad A_1 y(b) + B_1 y'(b) = 0. \tag{9.67b}$$

The functions $p(x)$, $p'(x)$, $q(x)$, $\varrho(x)$, and $f(x)$ are supposed to be continuous in the finite interval $a \le x \le b$. In the case of an infinite interval the results change considerably (see [9.5]). Furthermore, it is supposed that $p(x) > p_0 > 0$, $\varrho(x) > \varrho_0 > 0$. The quantity λ, a parameter of the differential equation, is a constant. For $f = 0$, it is called the *homogeneous boundary value problem* associated to the *inhomogeneous boundary value problem*.

Every second-order differential equation of the form

$$Ay'' + By' + Cy + \lambda R y = F \tag{9.67c}$$

can be reduced to the self-adjoint equation (9.67a) by multiplying it by p/A if in $[a, b]$, $A \neq 0$, and performing the following substitutions

$$p = \exp\left(\int \frac{R}{A}\,dx\right), \quad q = -\frac{pC}{A}, \quad \varrho = \frac{pR}{A}\,. \tag{9.67d}$$

To find a solution satisfying the inhomogeneous conditions

$$A_0 y(a) + B_0 y'(a) = C_0, \quad A_1 y(b) + B_1 y'(b) = C_1 \tag{9.67e}$$

one returns to the problem with homogeneous boundary conditions, but the right-hand side $f(x)$ changes and $y = z + u$ is substituted where u is an arbitrary twice differentiable function satisfying the inhomogeneous boundary conditions and z is a new unknown function satisfying the corresponding homogeneous conditions.

3. Sturm-Liouville Problem

For a given value of the parameter λ there are two cases:

1. Either the inhomogeneous boundary value problem has a unique solution for arbitrary $f(x)$, while the corresponding homogeneous problem has only the trivial, identically zero solution, or,

2. The corresponding homogeneous problem also has non-trivial, i.e., not identically zero solutions, but in this case the inhomogeneous problem does not have a solution for arbitrary right-hand side; and if a solution exists, it is not unique.

The values of the parameter λ, for which the second case occurs, i.e., the homogeneous problem has a non-trivial solution, are called the *eigenvalues of the boundary value problem*, the corresponding non-trivial solutions are called the *eigenfunctions*. The problem of determining the eigenvalues and eigenfunctions of a differential equation (9.67a) is called the *Sturm-Liouville problem*.

9.1.3.2 Fundamental Properties of Eigenfunctions and Eigenvalues

1. The eigenvalues of a boundary value problem form a monotone increasing sequence of real numbers

$$\lambda_0 < \lambda_1 < \lambda_2 < \cdots < \lambda_n < \cdots, \tag{9.68a}$$

tending to infinity.

2. The eigenfunction associated to the eigenvalue λ_n has exactly n roots in the interval $a < x < b$.
3. If $y(x)$ and $z(x)$ are two eigenfunctions belonging to the same eigenvalue λ, they differ only in a constant multiplier c, i.e.,

$$z(x) = cy(x). \tag{9.68b}$$

4. Two eigenfunctions $y_1(x)$ and $y_2(x)$, associated to different eigenvalues λ_1 and λ_2, are *orthogonal* to each other with the *weight function* $\varrho(x)$

$$\int_a^b y_1(x)\, y_2(x)\, \varrho(x)\, dx = 0. \tag{9.68c}$$

5. If in (9.67a) the coefficients $p(x)$ and $q(x)$ are replaced by $\tilde{p}(x) \geq p(x)$ and $\tilde{q}(x) \geq q(x)$, then the eigenvalues will not decrease, i.e., $\tilde{\lambda}_n \geq \lambda_n$, where $\tilde{\lambda}_n$ and λ_n are the n-th eigenvalues of the modified and the original equations respectively. But if the coefficient $\varrho(x)$ is replaced by $\tilde{\varrho}(x) \geq \varrho(x)$, then the eigenvalues will not increase, i.e., $\tilde{\lambda}_n \leq \lambda_n$. The n-th eigenvalue depends continuously on the coefficients of the equation, i.e., small changes in the coefficients will result in small variations of the n-th eigenvalue.
6. Reduction of the interval $[a, b]$ into a smaller one does not result in smaller eigenvalues.

9.1.3.3 Expansion in Eigenfunctions

1. Normalization of the Eigenfunction

For every λ_n an eigenfunction $\varphi_n(x)$ is chosen such that

$$\int_a^b [\varphi_n(x)]^2 \varrho(x)\, dx = 1. \tag{9.69a}$$

It is called a *normalized eigenfunction*.

2. Fourier Expansion

To every function $g(x)$ defined in the interval $[a, b]$, one can assign its *Fourier series*

$$g(x) \sim \sum_{n=0}^{\infty} c_n \varphi_n(x), \quad c_n = \int_a^b g(x)\, \varphi_n(x)\, \varrho(x)\, dx \tag{9.69b}$$

with the eigenfunctions of the corresponding boundary value problem, if the integrals in (9.69b) exist.

3. Expansion Theorem

If the function $g(x)$ has a continuous derivative and satisfies the boundary conditions of the given problem, then the Fourier series of $g(x)$ (in the eigenfunctions of this boundary value problem) is absolutely and uniformly convergent to $g(x)$.

4. Parseval Equation

If the integral on the left-hand side exists, then

$$\int_a^b [g(x)]^2 \varrho(x)\, dx = \sum_{n=0}^{\infty} {c_n}^2 \tag{9.69c}$$

is always valid. The Fourier series of the function $g(x)$ converges in this case to $g(x)$ in mean, that is

$$\lim_{N\to\infty} \int_a^b \left[g(x) - \sum_{n=0}^{N} c_n \varphi_n(x) \right]^2 \varrho(x)\, dx = 0. \tag{9.69d}$$

9.1.3.4 Singular Cases

Boundary value problems of the above type occur very often in solving problems of theoretical physics by the Fourier method, however at the endpoints of the interval $[a, b]$ some singularities of the differen-

tial equation may occur, e.g., $p(x)$ vanishes. At such singular points some restrictions are imposed on the solutions, e.g., continuity or being finite or unlimited growth with a bounded order. These conditions play the role of homogeneous boundary conditions (see 9.2.3.3, p. 591). In addition, often occurs the case where in certain boundary value problems homogeneous boundary conditions should be considered, such that they connect the values of the function or its derivative at different endpoints of the interval. Often occur the relations

$$y(a) = y(b), \quad p(a)y'(a) = p(b)y'(b), \tag{9.70}$$

which represent periodicity in the case of $p(a) = p(b)$. For such boundary value problems everything being introduced above remains valid, except statement (9.68b). For further discussion of this topic see [9.5].

9.2 Partial Differential Equations

9.2.1 First-Order Partial Differential Equations

9.2.1.1 Linear First-Order Partial Differential Equations

1. Linear and Quasilinear Partial Differential Equations

The equation

$$X_1 \frac{\partial z}{\partial x_1} + X_2 \frac{\partial z}{\partial x_2} + \cdots + X_n \frac{\partial z}{\partial x_n} = Y \tag{9.71a}$$

is called a *linear first-order partial differential equation.* Here z is an unknown function of the independent variables $x_1, \ldots, x_n$, and $X_1, \ldots, X_n, Y$ are given functions of these variables. If functions $X_1, \ldots, X_n, Y$ depend also on z, the equation is called a *quasilinear partial differential equation.* In the case of

$$Y \equiv 0, \tag{9.71b}$$

the equation is called homogeneous.

2. Solution of a Homogeneous Partial Linear Differential Equation

The solution of a homogeneous partial linear differential equation and the solution of the so-called *characteristic system*

$$\frac{dx_1}{X_1} = \frac{dx_2}{X_2} = \cdots = \frac{dx_n}{X_n} \tag{9.72a}$$

are equivalent. This system can be solved in two different ways:

1. Any x_k, for which $X_k \neq 0$, can be chosen as an independent variable, so the system is transformed into the form

$$\frac{dx_j}{dx_k} = \frac{X_j}{X_k} \quad (j = 1, \ldots, n). \tag{9.72b}$$

2. A more convenient way is to keep symmetry and to introduce a new variable t getting

$$\frac{dx_j}{dt} = X_j \quad (j = 1, 2, \ldots, n). \tag{9.72c}$$

Every first integral of the system (9.72a) is a solution of the homogeneous linear partial differential equation (9.72a,b), and conversely, every solution of (9.72a,b) is a first integral of (9.72a) (see 9.1.2.1, **2.**, p. 551). If the $n - 1$ first integrals

$$\varphi_i(x_1, \ldots, x_n) = 0 \quad (i = 1, 2, \ldots, n-1) \tag{9.72d}$$

are independent (see 9.1.2.3, **2.**, p. 553), then the general solution is

$$z = \Phi(\varphi_1, \ldots, \varphi_{n-1}). \tag{9.72e}$$

Here Φ is an arbitrary function of the $n - 1$ arguments φ_i and a general solution of the homogeneous linear differential equation.

3. Solution of Inhomogeneous Linear and Quasilinear Partial Differential Equations

To solve an inhomogeneous linear and quasilinear partial differential equation (9.71a) one can try to find the solution z in the implicit form $V(x_1, \ldots, x_n, z) = C$. The function V is a solution of the homogeneous linear differential equation with $n+1$ independent variables

$$X_1 \frac{\partial V}{\partial x_1} + X_2 \frac{\partial V}{\partial x_2} + \cdots + X_n \frac{\partial V}{\partial x_n} + Y \frac{\partial V}{\partial z} = 0, \tag{9.73a}$$

whose characteristic system

$$\frac{dx_1}{X_1} = \frac{dx_2}{X_2} = \cdots = \frac{dx_n}{X_n} = \frac{dz}{Y} \tag{9.73b}$$

is called the *characteristic system of the original equation* (9.71a).

4. Geometrical Representation and Characteristics of the System

In the case of the equation

$$P(x,y,z)\frac{\partial z}{\partial x} + Q(x,y,z)\frac{\partial z}{\partial y} = R(x,y,z) \tag{9.74a}$$

with two independent variables $x_1 = x$ and $x_2 = y$, a solution $z = f(x,y)$ is a surface in x, y, z space, and it is called the *integral surface* of the differential equation. Equation (9.74a) means that at every point of the integral surface $z = f(x,y)$ the normal vector $\left(\frac{\partial z}{\partial x}, \frac{\partial z}{\partial y}, -1\right)$ is orthogonal to the vector (P, Q, R) given at that point. Here the system (9.73b) has the form

$$\frac{dx}{P(x,y,z)} = \frac{dy}{Q(x,y,z)} = \frac{dz}{R(x,y,z)}\,. \tag{9.74b}$$

It follows (see 13.1.3.5, p. 708) that the *integral curves of this system*, the so-called *characteristics*, are tangent to the vector (P, Q, R). Therefore, a characteristic having a common point with the integral surface $z = f(x,y)$ lies completely on this surface. Since the conditions for the existence theorem 13.1.3.5, **1.**, p. 551 hold, there is an integral curve of the characteristic system passing through every point of space, so the integral surface consists of characteristics.

5. Cauchy Problem

There are given n functions of $n-1$ independent variables $t_1, t_2, \ldots, t_{n-1}$:

$$x_1 = x_1(t_1, t_2, \ldots, t_{n-1}),\; x_2 = x_2(t_1, t_2, \ldots, t_{n-1}), \ldots,\; x_n = x_n(t_1, t_2, \ldots, t_{n-1}). \tag{9.75a}$$

The Cauchy problem for the differential equation (9.71a) is to find a solution

$$z = \varphi(x_1, x_2, \ldots, x_n) \tag{9.75b}$$

such that if one substitutes (9.75a), the result is a previously given function $\psi(t_1, t_2, \ldots, t_{n-1})$:

$$\varphi[x_1(t_1, t_2, \ldots, t_{n-1}),\; x_2(t_1, t_2, \ldots, t_{n-1}), \ldots,\; x_n(t_1, t_2, \ldots, t_{n-1})] = \psi(t_1, t_2, \ldots, t_{n-1}). \tag{9.75c}$$

In the case of two independent variables, the problem reduces to find an integral surface passing through the given curve. If this curve has a tangent depending continuously on a point and it is not tangent to the characteristics at any point, then the Cauchy problem has a unique solution in a certain neighborhood of this curve. Here the integral surface consists of the set of all characteristics intersecting the given curve. For more mathematical discussion on theorems about the existence of the solution of the Cauchy problem see [9.15].

■ **A:** For the linear first-order inhomogeneous partial differential equation $(mz - ny)\frac{\partial z}{\partial x} + (nx - lz)\frac{\partial z}{\partial y} = ly - mx$ (l, m, n are constants), the equations of the characteristics are $\frac{dx}{mz - ny} = \frac{dy}{nx - lz} =$

$\dfrac{dz}{ly - mx}$. The integrals of this system are $lx + my + nz = C_1$, $x^2 + y^2 + z^2 = C_2$. One gets circles as characteristics, whose centers are on a line passing through the origin, and this line has direction cosines proportional to l, m, n. The integral surfaces are rotation surfaces with this line as an axis.

■ **B:** Determine the integral surface of the first-order linear inhomogeneous differential equation $\dfrac{\partial z}{\partial x} + \dfrac{\partial z}{\partial y} = z$, which passes through the curve $x = 0$, $z = \varphi(y)$. The equations of characteristics are $\dfrac{dx}{1} = \dfrac{dy}{1} = \dfrac{dz}{z}$. The characteristics passing through the point (x_0, y_0, z_0) are $y = x - x_0 + y_0$, $z = z_0 e^{x-x_0}$. A parametric representation of the required integral surface is $y = x + y_0$, $z = e^x \varphi(y_0)$, if we substitute $x_0 = 0$, $z_0 = \varphi(y_0)$. The elimination of y_0 results in $z = e^x \varphi(y - x)$.

9.2.1.2 Non-Linear First-Order Partial Differential Equations

1. General Form of First-Order Partial Differential Equation

is the implicit equation

$$F\left(x_1, \ldots, x_n, z, \frac{\partial z}{\partial x_1}, \ldots, \frac{\partial z}{\partial x_n}\right) = 0. \tag{9.76a}$$

1. Complete Integral is the solution

$$z = \varphi(x_1, \ldots, x_n; a_1, \ldots, a_n), \tag{9.76b}$$

depending on n parameters $a_1, \ldots, a_n$ if at the considered values of $x_1, \ldots, x_n, z$ the functional determinant (or Jacobian determinant, see 2.18.2.6, **3.**, p. 123) is non-zero:

$$\frac{\partial(\varphi_{x_1}, \ldots, \varphi_{x_n})}{\partial(a_1, \ldots, a_n)} \neq 0. \tag{9.76c}$$

2. Characteristic Strip The solution of (9.76a) is reduced to the solution of the characteristic system

$$\frac{dx_1}{P_1} = \cdots = \frac{dx_n}{P_n} = \frac{dz}{p_1P_1 + \cdots + p_nP_n} = \frac{-dp_1}{X_1 + p_1Z} = \cdots = \frac{-dp_n}{X_n + p_nZ} \tag{9.76d}$$

with

$$Z = \frac{\partial F}{\partial z}, \quad X_i = \frac{\partial F}{\partial x_i}, \quad p_i = \frac{\partial z}{\partial x_i}, \quad P_i = \frac{\partial F}{\partial p_i} \quad (i = 1, \ldots, n). \tag{9.76e}$$

The solutions of the characteristic system satisfying the additional condition

$$F(x_1, \ldots, x_n, z, p_1, \ldots, p_n) = 0 \tag{9.76f}$$

are called the *characteristic strips*.

2. Canonical Systems of Differential Equations

Sometimes it is more convenient to consider an equation not involving explicitly the unknown function z. Such an equation can be obtained by introducing an additional independent variable $x_{n+1} = z$ and an unknown function $V(x_1, \ldots, x_n, x_{n+1})$, which defines the function $z(x_1, x_2, \ldots, x_n)$ with the equation

$$V(x_1, \ldots, x_n, z) = C. \tag{9.77a}$$

At the same time, one substitutes the functions $-\dfrac{\partial V}{\partial x_i} \Big/ \dfrac{\partial V}{\partial x_{n+1}}$ $(i = 1, \ldots, n)$ for $\dfrac{\partial z}{\partial x_i}$ in (9.76a). Then one solves the differential equation (9.76a) for an arbitrary partial derivative of the function V.

The corresponding independent variable will be denoted by x after a suitable renumbering of the other variables. Finally, one gets the equation (9.76a) in the form

$$p + H(x_1, \dots, x_n, x, p_1, \dots, p_n) = 0, \ \ p = \frac{\partial V}{\partial x}, \ \ p_i = \frac{\partial V}{\partial x_i} \quad (i = 1, \dots, n). \tag{9.77b}$$

The system of characteristic differential equations is transformed into the system

$$\frac{dx_i}{dx} = \frac{\partial H}{\partial p_i}, \ \ \frac{dp_i}{dx} = -\frac{\partial H}{\partial x_i} \quad (i = 1, \dots, n) \quad \text{and} \tag{9.77c}$$

$$\frac{dV}{dx} = p_1 \frac{\partial H}{\partial p_1} + \cdots + p_n \frac{\partial H}{\partial p_n} - H, \quad \frac{dp}{dx} = -\frac{\partial H}{\partial x}. \tag{9.77d}$$

Equations (9.77c) represent a system of $2n$ ordinary differential equations, which corresponds to an arbitrary function $H(x_1, \dots, x_n, x, p_1, \dots, p_n)$ with $2n + 1$ variables. It is called a *canonical system* or a *normal system* of differential equations.

Many problems of mechanics and theoretical physics lead to equations of this form. Knowing a complete integral

$$V = \varphi(x_1, \dots, x_n, x, a_1, \dots, a_n) + a \tag{9.77e}$$

of the equation (9.77b) one can find the general solution of the canonical system (9.77c), since the equations $\frac{\partial \varphi}{\partial a_i} = b_i$, $\frac{\partial \varphi}{\partial x_i} = p_i$ $(i = 1, 2, \dots, n)$ with $2n$ arbitrary parameters a_i and b_i determine a $2n$-parameter solution of the canonical system (9.77c).

3. Clairaut Differential Equation

If the given differential equation can be transformed into the form

$$z = x_1 p_1 + x_2 p_2 + \cdots + x_n p_n + f(p_1, p_2, \dots, p_n), \ \ p_i = \frac{\partial z}{\partial x_i} \quad (i = 1, \dots, n), \tag{9.78a}$$

it is called a Clairaut differential equation. The determination of the complete integral is particularly simple, because a complete integral with the arbitrary parameters $a_1, a_2, \dots, a_n$ is

$$z = a_1 x_1 + a_2 x_2 + \cdots + a_n x_n + f(a_1, a_2, \dots, a_n). \tag{9.78b}$$

■ **Two-Body Problem with Hamilton Function:** Consider two particles moving in a plane under their mutual gravitational attraction according to the Newton field (see also 13.4.3.2, p. 728). Choosing the origin as the initial position of one of the particles, the equations of motion have the form

$$\frac{d^2x}{dt^2} = \frac{\partial V}{\partial x}, \ \ \frac{d^2y}{dt^2} = \frac{\partial V}{\partial y}; \quad V = \frac{k^2}{\sqrt{x^2 + y^2}}. \tag{9.79a}$$

Introducing the Hamiltonian function

$$H = \frac{1}{2}(p^2 + q^2) - \frac{k^2}{\sqrt{x^2 + y^2}}, \tag{9.79b}$$

the system (9.79a) is transformed into the normal system (into the system of canonical differential equations)

$$\frac{dx}{dt} = \frac{\partial H}{\partial p}, \ \ \frac{dy}{dt} = \frac{\partial H}{\partial q}, \ \ \frac{dp}{dt} = -\frac{\partial H}{\partial x}, \ \ \frac{dq}{dt} = -\frac{\partial H}{\partial y} \tag{9.79c}$$

with variables

$$x, y, p = \frac{dx}{dt}, \ \ q = \frac{dy}{dt}. \tag{9.79d}$$

Now, the partial differential equation has the form

$$\frac{\partial z}{\partial t} + \frac{1}{2}\left[\left(\frac{\partial z}{\partial x}\right)^2 + \left(\frac{\partial z}{\partial y}\right)^2\right] - \frac{k^2}{\sqrt{x^2+y^2}} = 0. \tag{9.79e}$$

Introducing the polar coordinates ρ, φ in (9.79e) one obtains a new differential equation having the solution

$$z = -at - b\varphi + c - \int_{\rho_0}^{\rho} \sqrt{2a + \frac{2k^2}{r} - \frac{b^2}{r^2}}\, dr \tag{9.79f}$$

with the parameters a, b, c. The general solution of the system (9.79c) follows from the equations

$$\frac{\partial z}{\partial a} = -t_0, \quad \frac{\partial z}{\partial b} = -\varphi_0. \tag{9.79g}$$

4. First-Order Differential Equation in Two Independent Variables

For $x_1 = x$, $x_2 = y$, $p_1 = p$, $p_2 = q$ the characteristic strip (see 9.2.1.2, **1.**, p. 573) can be geometrically interpreted as a curve at every point (x, y, z) of which a plane $p(\xi - x) + q(\eta - y) = \zeta - z$ being tangent to the curve is prescribed. So, the problem of finding an integral surface of the equation

$$F\left(x, y, z, \frac{\partial z}{\partial x}, \frac{\partial z}{\partial y}\right) = 0 \tag{9.80}$$

passing through a given curve, i.e., to solve the Cauchy problem (see 9.2.1.1, **5.**, p. 572), is transformed into another problem: To find the characteristic strips passing through the points of the initial curve such that the corresponding tangent plane to each strip is tangent to that curve. One gets the values p and q at the points of the initial curve from the equations $F(x, y, z, p, q) = 0$ and $pdx + qdy = dz$. There can be several solutions in the case of non-linear differential equations.

Therefore, under the formulation of the Cauchy problem, in order to obtain a unique solution one can assume two continuous functions p and q satisfying the above relations along the initial curve.
For the existence of solutions of the Cauchy problem see [9.15].

■ For the partial differential equation $pq = 1$ and the initial curve $y = x^3$, $z = 2x^2$, one can choose $p = x$ and $q = 1/x$ along the curve. The characteristic system has the form

$$\frac{dx}{dt} = q, \quad \frac{dy}{dt} = p, \quad \frac{dz}{dt} = 2pq, \quad \frac{dp}{dt} = 0, \quad \frac{dq}{dt} = 0.$$

The characteristic strip with initial values x_0, y_0, z_0, p_0 and q_0 for $t = 0$ satisfies the equations $x = x_0 + q_0 t$, $y = y_0 + p_0 t$, $z = 2p_0 q_0 t + z_0$, $p = p_0, q = q_0$. For the case of $p_0 = x_0$, $q_0 = 1/x_0$ the equation of the curve belonging to the characteristic strip that passes through the point (x_0, y_0, z_0) of the initial curve is

$$x = x_0 + \frac{t}{x_0}, \quad y = {x_0}^3 + tx_0, \quad z = 2t + 2{x_0}^2.$$

Eliminating the parameters x_0 and t gives $z^2 = 4xy$. For other chosen values of p and q along the initial curve one can get different solutions.

Remark: The envelope of a one-parameter family of integral surfaces is also an integral surface. Considering this fact one can solve the Cauchy problem with a complete integral. One finds a one-parameter family of solutions tangent to the planes given at the points of the initial curve. Then one determines the envelope of this family.

■ Determine the integral surface for the Clairaut differential equation $z - px - qy + pq = 0$ passing through the curve $y = x$, $z = x^2$. The complete integral of the differential equation is $z = ax + by - ab$. Since along the initial curve $p = q = x$, one determines the one-parameter family of integral surfaces by the condition $a = b$. When the envelope of this family is found then one gets $z = \frac{1}{4}(x+y)^2$.

5. Linear First-Order Partial Differential Equations in Total Differentials

Equations of this kind have the form

$$dz = f_1 dx_1 + f_2 dx_2 + \cdots + f_n dx_n, \tag{9.81a}$$

where $f_1, f_2, \ldots, f_n$ are given functions of the variables $x_1, x_2, \ldots, x_n, z$. The equation is called a *completely integrable* or *exact differential equation* when there exists a unique relation between $x_1, x_2, \ldots, x_n$, z with one arbitrary constant, which leads to equation (9.81a). Then there exists a unique solution $z = z(x_1, x_2, \ldots, x_n)$ of (9.81a), which has a given value z_0 for the initial values ${x_1}^0, \ldots, {x_n}^0$ of the independent variables. Therefore, for $n = 2$, $x_1 = x$, $x_2 = y$ a unique integral surface passes through every point of space.

The differential equation (9.81a) is *completely integrable* if and only if the $\dfrac{n(n-1)}{2}$ equalities

$$\frac{\partial f_i}{\partial x_k} + f_k \frac{\partial f_i}{\partial z} = \frac{\partial f_k}{\partial x_i} + f_i \frac{\partial f_k}{\partial z} \quad (i, k = 1, \ldots, n) \tag{9.81b}$$

in all variables $x_1, x_2, \ldots, x_n, z$ are identically satisfied.
If the differential equation is given in symmetric form

$$f_1 dx_1 + \cdots + f_n dx_n = 0, \tag{9.81c}$$

then the condition for complete integrability is

$$f_i \left(\frac{\partial f_k}{\partial x_j} - \frac{\partial f_j}{\partial x_k} \right) + f_j \left(\frac{\partial f_i}{\partial x_k} - \frac{\partial f_k}{\partial x_i} \right) + f_k \left(\frac{\partial f_j}{\partial x_i} - \frac{\partial f_i}{\partial x_j} \right) = 0 \tag{9.81d}$$

for all possible combinations of the indices i, j, k. If the equation is completely integrable, then the solution of the differential equation (9.81a) can be reduced to the solution of an ordinary differential equation with $n-1$ parameters.

9.2.2 Linear Second-Order Partial Differential Equations

9.2.2.1 Classification and Properties of Second-Order Differential Equations with Two Independent Variables

1. General Form

of a linear second-order partial differential equation with two independent variables x, y and an unknown function u is an equation in the form

$$A \frac{\partial^2 u}{\partial x^2} + 2B \frac{\partial^2 u}{\partial x \partial y} + C \frac{\partial^2 u}{\partial y^2} + a \frac{\partial u}{\partial x} + b \frac{\partial u}{\partial y} + cu = f, \tag{9.82a}$$

where the coefficients A, B, C, a, b, c and f on the right-hand side are known functions of x and y.
The form of the solution of this differential equation depends on the sign of the *discriminant*

$$\delta = AC - B^2 \tag{9.82b}$$

in a considered domain. The following cases should be distinguished.

1. $\delta < 0$: Hyperbolic type.

2. $\delta = 0$: Parabolic type.

3. $\delta > 0$: Elliptic type.

4. δ changes its sign: **Mixed type.**

An important property of the discriminant δ is that its sign is invariant with respect to arbitrary transformation of the independent variables, e.g., to introduction new coordinates in the x, y plane. Therefore, the type of the differential equation is invariant with respect to the choice of the independent variables.

2. Characteristics

of linear second-order partial differential equations are the integral curves of the differential equation

$$Ady^2 - 2Bdxdy + Cdx^2 = 0 \quad \text{or} \quad \frac{dy}{dx} = \frac{B \pm \sqrt{-\delta}}{A}. \tag{9.83}$$

For the characteristics of the above three types of differential equations the following statements are valid:

1. Hyperbolic type: There exist two families of real characteristics.

2. Parabolic type: There exists only one family of real characteristics.

3. Elliptic type: There exists no real characteristic.

4. A differential equation obtained by coordinate transformation from (9.82a) has the same characteristics as (9.82a).

5. If a family of characteristics coincides with a family of coordinate lines, then the term with the second derivative of the unknown function with respect to the corresponding independent variable is missing in (9.82a). In the case of a parabolic differential equation, the mixed derivative term is also missing.

3. Normal Form or Canonical Form

One has the following possibilities to transform (9.82a) into the normal form of linear second-order partial differential equations.

1. Transformation into Normal Form: The differential equation (9.82a) can be transformed into normal form by introducing the new independent variables

$$\xi = \varphi(x, y) \quad \text{and} \quad \eta = \psi(x, y), \tag{9.84a}$$

which according to the sign of the discriminant (9.82b) belongs to one of the three considered types:

$$\frac{\partial^2 u}{\partial \xi^2} - \frac{\partial^2 u}{\partial \eta^2} + \cdots = 0, \quad \delta < 0, \quad \text{hyperbolic type;} \tag{9.84b}$$

$$\frac{\partial^2 u}{\partial \eta^2} + \cdots = 0, \quad \delta = 0, \quad \text{parabolic type;} \tag{9.84c}$$

$$\frac{\partial^2 u}{\partial \xi^2} + \frac{\partial^2 u}{\partial \eta^2} + \cdots = 0, \quad \delta > 0, \quad \text{elliptic type.} \tag{9.84d}$$

The terms not containing second-order partial derivatives of the unknown function are denoted by dots.

2. Reduction of a Hyperbolic Type Equation to Canonical Form (9.84b)**:** If, in the hyperbolic case, one chooses two families of characteristics as the coordinate lines of the new coordinate system (9.84a), i.e., if substituting $\xi_1 = \varphi(x, y)$, $\eta_1 = \psi(x, y)$, where $\varphi(x, y) =$ constant, $\psi(x, y) =$ constant are the equations of the characteristics, then (9.82a) becomes the form

$$\frac{\partial^2 u}{\partial \xi_1 \partial \eta_1} + \cdots = 0. \tag{9.84e}$$

This form is also called the *canonical form of a hyperbolic type differential equation*. From here one gets the canonical form (9.84b) by the substitution

$$\xi = \xi_1 + \eta_1, \quad \eta = \xi_1 - \eta_1. \tag{9.84f}$$

3. Reduction of a Parabolic Type Equation to Canonical Form (9.84c)**:** The only family of characteristics given in this case is selected for the family $\xi =$ const, where an arbitrary function of x and y can be chosen for η, which must not be dependent on ξ.

4. Reduction of an Elliptic Type Equation to Canonical Form (9.84d)**:** If the coefficients $A(x, y)$, $B(x, y)$, $C(x, y)$ are analytic functions (see 14.1.2.1, p. 732) in the elliptic case, then the characteristics define two complex conjugate families of curves $\varphi(x, y) =$ constant, $\psi(x, y) =$ constant. By substituting $\xi = \varphi + \psi$ and $\eta = \mathrm{i}(\varphi - \psi)$, the equation becomes of the form (9.84d).

4. Generalized Form

Every statement for the classification and reduction to canonical forms remains valid for equations given in a more general form

$$A(x,y)\frac{\partial^2 u}{\partial x^2} + 2B(x,y)\frac{\partial^2 u}{\partial x \partial y} + C(x,y)\frac{\partial^2 u}{\partial y^2} + F\left(x,y,u,\frac{\partial u}{\partial x},\frac{\partial u}{\partial y}\right) = 0, \tag{9.85}$$

where F is a non-linear function of the unknown function u and its first-order partial derivatives $\partial u/\partial x$ and $\partial u/\partial y$, in contrast to (9.82a).

9.2.2.2 Classification and Properties of Linear Second-Order Differential Equations with More than Two Independent Variables

1. General Form

A differential equation of this kind for $u = u(x_1, x_2, \ldots, x_n)$ has the form

$$\sum_{i,k} a_{ik} \frac{\partial^2 u}{\partial x_i \partial x_k} + \cdots = 0, \tag{9.86}$$

where a_{ik} are given functions of the independent variables and the dots in (9.86) mean terms not containing second-order derivatives of the unknown function.
In general, the differential equation (9.86) cannot be reduced to a simple canonical form by transforming the independent variables. However, there is an important classification, similar to the one introduced above in 9.2.2.1, p. 576 (see [9.5]).

2. Linear Second-Order Partial Differential Equations with Constant Coefficients

If all coefficients a_{ik} in (9.86) are constants, then the equation can be reduced by a linear homogeneous transformation of the independent variables into a simpler canonical form

$$\sum_{i} \kappa_i \frac{\partial^2 u}{\partial {x_i}^2} + \cdots = 0, \tag{9.87}$$

where the coefficients κ_i are ± 1 or 0. Several characteristic cases have to be distinguished.

1. Elliptic Differential Equation If all coefficients κ_i are different from zero, and they have the same sign, then it is the case of an *elliptic differential equation.*

2. Hyperbolic and Ultra-Hyperbolic Differential Equation If all coefficients κ_i are different from zero, but one has a sign different from the other's, then it is the case of a *hyperbolic differential equation.* If both types of signs occur at least twice, then it is an *ultra-hyperbolic differential equation.*

3. Parabolic Differential Equation If one of the coefficients κ_i is equal to zero, the others are different from zero and they have the same sign, then it is the case of a *parabolic differential equation.*

4. Simple Case for Elliptic and Hyperbolic Differential Equations If not only the coefficients of the second order derivatives of the unknown function are constants, but also those of the first order derivatives, then it is possible to eliminate the terms of the first order derivatives, for which $\kappa_i \neq 0$, by substitution. For this purpose is

$$u = v \exp\left(-\frac{1}{2}\sum \frac{b_k}{\kappa_k} x_k\right), \tag{9.88}$$

substituted where b_k is the coefficient of $\dfrac{\partial u}{\partial x_k}$ in (9.87) and the summation is performed for all $\kappa_i \neq 0$.
In this way, every elliptic and hyperbolic differential equation with constant coefficients can be reduced to a simple form:

a) Elliptic Case: $\Delta v + kv = g.$ (9.89) **b) Hyperbolic Case:** $\dfrac{\partial^2 v}{\partial t^2} - \Delta v + kv = g.$ (9.90)

Here Δ denotes the Laplace operator (see 13.2.6.5, p. 716).

9.2.2.3 Integration Methods for Linear Second-Order Partial Differential Equations

1. Method of Separation of Variables

Certain solutions of several differential equations of physics can be determined by special substitutions, and although these are not general solutions, one gets a family of solutions depending on arbitrary parameters. Linear differential equations, especially those of second order, can often be solved if looking for a solution in the *form of a product*

$$u(x_1, \dots, x_n) = \varphi_1(x_1)\varphi_2(x_2)\dots\varphi_n(x_n). \tag{9.91}$$

Next, one tries to separate the functions $\varphi_k(x_k)$, i.e., for each of them one wants to determine an ordinary differential equation containing only one variable x_k. This *separation of variables* is successful in many cases when the trial solution in the form of a product (9.91) is substituted into the given differential equation. In order to guarantee that the solution of the original equation satisfies the required homogeneous boundary conditions, it may appear to be sufficient that some of functions $\varphi_1(x_1)$, $\varphi_2(x_2), \dots, \varphi_n(x_n)$ satisfy certain boundary conditions.

By means of summation, differentiation and integration, new solutions can be acquired from the obtained ones; the parameters should be chosen so that the remaining boundary and initial conditions are satisfied (see examples).

Finally, don't forget that the solutions obtained in this way, often infinite series and improper integrals, are only *formal solutions*. That is, one has to check whether the solution makes a physical sense, e.g., whether it is convergent, satisfies the original differential equation and the boundary conditions, whether it is differentiable termwise and whether the limit at the boundary exists.

The infinite series and improper integrals in the examples of this paragraph are convergent if the functions defining the boundary conditions satisfy the required conditions, e.g., the continuity assumption for the second derivatives in the first and the second examples.

■ **A: Equation of the Vibrating String** is a linear second-order partial differential equation of hyperbolic type

$$\frac{\partial^2 u}{\partial t^2} = a^2 \frac{\partial^2 u}{\partial x^2}. \tag{9.92a}$$

It describes the vibration of a spanned string. The boundary and the initial conditions are:

$$u\Big|_{t=0} = f(x), \quad \frac{\partial u}{\partial t}\Big|_{t=0} = \varphi(x), \quad u|_{x=0} = 0, \quad u|_{x=l} = 0. \tag{9.92b}$$

Seeking a solution in the form

$$u = X(x)T(t), \tag{9.92c}$$

and after substituting it into the given equation (9.92a) follows

$$\frac{T''}{a^2 T} = \frac{X''}{X}. \tag{9.92d}$$

The variables are separated, the right side depends on only x and the left side depends on only t, so each of them is a constant quantity. The constant must be negative, otherwise the boundary conditions cannot be satisfied, i.e., non-negative values give the trivial solution $u(x,t) = 0$. This negative constant is denoted by $-\lambda^2$. The result is an ordinary linear second-order differential equation with constant coefficients for both variables. For the general solution see 9.1.2.4, p. 555. The results are the linear differential equations

$$X'' + \lambda^2 X = 0, \tag{9.92e}$$

and

$$T'' + a^2\lambda^2 T = 0. \tag{9.92f}$$

From the boundary conditions follows $X(0) = X(l) = 0$. Hence $X(x)$ is an eigenfunction of the

Sturm-Liouville boundary value problem and λ^2 is the corresponding eigenvalue (see 9.1.3.1, **3.**, p. 569). Solving the differential equation (9.92e) for X with the corresponding boundary conditions one gets

$$X(x) = C \sin \lambda x \quad \text{with} \quad \sin \lambda l = 0, \quad \text{i.e., with} \quad \lambda = \frac{n\pi}{l} = \lambda_n \quad (n = 1, 2, \ldots). \tag{9.92g}$$

Solving equation (9.92f) for T yields a particular solution of the original differential equation (9.92a) for every eigenvalue λ_n:

$$u_n(x,t) = \left(a_n \cos \frac{na\pi}{l} t + b_n \sin \frac{na\pi}{l} t\right) \sin \frac{n\pi}{l} x. \tag{9.92h}$$

Requiring that for $t = 0$,

$$u\Big|_{t=0} = \sum_{n=1}^{\infty} u_n(x,0) \text{ is equal to f(x)}, \quad \text{and} \tag{9.92i}$$

$$\frac{\partial u}{\partial t}\Big|_{t=0} = \sum_{n=1}^{\infty} \frac{\partial u_n}{\partial t}(x,0) \text{ is equal to } \varphi(x), \tag{9.92j}$$

one gets with a Fourier series expansion in sines (see 7.4.1.1, **1.**, p. 474)

$$a_n = \frac{2}{l}\int_0^l f(x) \sin \frac{n\pi x}{l} dx, \quad b_n = \frac{2}{na\pi}\int_0^l \varphi(x) \sin \frac{n\pi x}{l} dx. \tag{9.92k}$$

■ **B: Equation of Longitudinal Vibration of a Bar** is a linear second-order partial differential equation of hyperbolic type, which describes the longitudinal vibration of a bar with one end free and a constant force p affecting the fixed end. Here is to solve the same differential equation as in ■ **A** (p. 579), i.e.,

$$\frac{\partial^2 u}{\partial t^2} = a^2 \frac{\partial^2 u}{\partial x^2}, \tag{9.93a}$$

with the same initial but different boundary conditions:

$$u\Big|_{t=0} = f(x), \quad \frac{\partial u}{\partial t}\Big|_{t=0} = \varphi(x), \quad (9.93\text{b}) \qquad \frac{\partial u}{\partial x}\Big|_{x=0} = 0 \quad \text{(free end)}, \tag{9.93c}$$

$$\frac{\partial u}{\partial x}\Big|_{x=l} = kp. \tag{9.93d}$$

The conditions (9.93c,d) can be replaced by the homogeneous conditions

$$\frac{\partial z}{\partial x}\Big|_{x=0} = \frac{\partial z}{\partial x}\Big|_{x=l} = 0 \tag{9.93e}$$

where instead of u is introduced a new unknown function

$$z = u - \frac{kpx^2}{2l}. \tag{9.93f}$$

The differential equation becomes inhomogeneous:

$$\frac{\partial^2 z}{\partial t^2} = a^2 \frac{\partial^2 z}{\partial x^2} + \frac{a^2 kp}{l}. \tag{9.93g}$$

Looking for the solution in the form $z = v + w$, where v satisfies the homogeneous differential equation with the initial and boundary conditions for z, i.e.,

$$z\Big|_{t=0} = f(x) - \frac{kpx^2}{2}, \quad \frac{\partial z}{\partial t}\Big|_{t=0} = \varphi(x), \tag{9.93h}$$

and w satisfies the inhomogeneous differential equation with zero initial and boundary conditions. This gives $w = \dfrac{ka^2pt^2}{2l}$. Substituting the product form of the unknown function $v(x,t)$ into the differential equation (9.93a)

$$v = X(x)T(t) \tag{9.93i}$$

gives the separated ordinary differential equations as in ■ **A** (p. 579)

$$\frac{X''}{X} = \frac{T''}{a^2T} = -\lambda^2. \tag{9.93j}$$

Integrating the differential equation for X with the boundary conditions $X'(0) = X'(l) = 0$ one finds the eigenfunctions

$$X_n = \cos\frac{n\pi x}{l} \tag{9.93k}$$

and the corresponding eigenvalues

$$\lambda_n{}^2 = \frac{n^2\pi^2}{l^2} \quad (n = 0, 1, 2, \ldots). \tag{9.93l}$$

Proceeding as in ■ **A** (p. 579) one finally obtains

$$u = \frac{ka^2pt^2}{2l} + \frac{kpx^2}{2l} + a_0 + \frac{a\pi}{l}b_0t + \sum_{n=1}^{\infty}\left(a_n\cos\frac{an\pi t}{l} + \frac{b_n}{n}\sin\frac{an\pi t}{l}\right)\cos\frac{n\pi x}{l}, \tag{9.93m}$$

where a_n and b_n $(n = 0, 1, 2, \ldots)$ are the coefficients of the Fourier series expansion in cosines of the functions $f(x) - \dfrac{kpx^2}{2}$ and $\dfrac{l}{a\pi}\varphi(x)$ in the interval $(0, l)$ (see 7.4.1.1, **1.**, p. 474).

■ **C: Equation of a Vibrating Round Membrane** fixed along the boundary:
The differential equation is linear, partial and it is of hyperbolic type. It has the form in Cartesian and in polar coordinates (see 3.5.3.2,**3.**, p. 211)

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = \frac{1}{a^2}\frac{\partial^2 u}{\partial t^2}, \tag{9.94a}$$

$$\frac{\partial^2 u}{\partial \rho^2} + \frac{1}{\rho}\frac{\partial u}{\partial \rho} + \frac{1}{\rho^2}\frac{\partial^2 u}{\partial \varphi^2} = \frac{1}{a^2}\frac{\partial^2 u}{\partial t^2}. \tag{9.94b}$$

The initial and boundary conditions are

$$u|_{t=0} = f(\rho,\varphi), \tag{9.94c}$$

$$\left.\frac{\partial u}{\partial t}\right|_{t=0} = F(\rho,\varphi), \tag{9.94d}$$

$$u|_{\rho=R} = 0. \tag{9.94e}$$

The substitution of the product form

$$u = U(\rho)\Phi(\varphi)T(t) \tag{9.94f}$$

with three variables into the differential equation in polar coordinates yields

$$\frac{U''}{U} + \frac{U'}{\rho U} + \frac{\Phi''}{\rho^2\Phi} = \frac{1}{a^2}\frac{T''}{T} = -\lambda^2. \tag{9.94g}$$

Three ordinary differential equations are obtained for the separated variables analogously to examples **A** (p. 579) and **B** (p. 580):

$$T'' + a^2\lambda^2 T = 0, \tag{9.94h}$$

$$\frac{\rho^2U'' + \rho U'}{U} + \lambda^2\rho^2 = -\frac{\Phi''}{\Phi} = \nu^2, \tag{9.94i}$$

$$\Phi'' + \nu^2\Phi = 0. \tag{9.94j}$$

From the conditions $\Phi(0) = \Phi(2\pi)$, $\Phi'(0) = \Phi'(2\pi)$ it follows that:

$$\Phi(\varphi) = a_n\cos n\varphi + b_n\sin n\varphi, \quad \nu^2 = n^2 \quad (n = 0, 1, 2, \ldots). \tag{9.94k}$$

U and λ will be determined from the equations $[\rho U']' - \dfrac{n^2}{\rho} U = -\lambda^2 \rho U$ and $U(R) = 0$. Considering the obvious condition of boundedness of $U(\rho)$ at $\rho = 0$ and substituting $\lambda\rho = z$ gives

$$z^2 U'' + zU' + (z^2 - n^2)U = 0, \quad \text{i.e.,} \quad U(\rho) = J_n(z) = J_n\left(\mu \frac{\rho}{R}\right), \tag{9.94l}$$

where J_n are the Bessel functions (see 9.1.2.6, **2.**, p. 562) with $\lambda = \dfrac{\mu}{R}$ and $J_n(\mu) = 0$. The system of functions

$$U_{nk}(\rho) = J_n\left(\mu_{nk} \frac{\rho}{R}\right) \quad (k = 1, 2, \ldots) \tag{9.94m}$$

with μ_{nk} as the k-th positive root of the function $J_n(z)$ is a complete system of eigenfunctions of the self-adjoint Sturm-Liouville problem which are orthogonal with the weight function ρ.
The solution of the problem can have the form of a double series:

$$U = \sum_{n=0}^{\infty} \sum_{k=1}^{\infty} \left[(a_{nk} \cos n\varphi + b_{nk} \sin n\varphi) \cos \frac{a\mu_{nk} t}{R} + (c_{nk} \cos n\varphi + d_{nk} \sin n\varphi) \sin \frac{a\mu_{nk} t}{R} \right] J_n\left(\mu_{nk} \frac{\rho}{R}\right). \tag{9.94n}$$

From the initial conditions at $t = 0$ one obtains

$$f(\rho, \varphi) = \sum_{n=0}^{\infty} \sum_{k=1}^{\infty} (a_{nk} \cos n\varphi + b_{nk} \sin n\varphi) J_n\left(\mu_{nk} \frac{\rho}{R}\right), \tag{9.94o}$$

$$F(\rho, \varphi) = \sum_{n=0}^{\infty} \sum_{k=1}^{\infty} \frac{a\mu_{nk}}{R} (c_{nk} \cos n\varphi + d_{nk} \sin n\varphi) J_n\left(\mu_{nk} \frac{\rho}{R}\right), \tag{9.94p}$$

where

$$a_{nk} = \frac{2}{\pi R^2 J_{n-1}^2(\mu_{nk})} \int_0^{2\pi} d\varphi \int_0^R f(\rho, \varphi) \cos n\varphi J_n\left(\mu_{nk} \frac{\rho}{R}\right) \rho \, d\rho, \tag{9.94q}$$

$$b_{nk} = \frac{2}{\pi R^2 J_{n-1}^2(\mu_{nk})} \int_0^{2\pi} d\varphi \int_0^R f(\rho, \varphi) \sin n\varphi J_n\left(\mu_{nk} \frac{\rho}{R}\right) \rho \, d\rho. \tag{9.94r}$$

In the case of $n = 0$, the numerator 2 should be changed to 1. To determine the coefficients c_{nk} and d_{nk} the function $f(\rho, \varphi)$ is replaced by $F(\rho, \varphi)$ in the formulas for a_{nk} and b_{nk} and finely it is multiplied by $\dfrac{R}{a\mu_{nk}}$.

■ **D: Dirichlet Problem** (see 13.5.1, p. 729) for the rectangle $0 \le x \le a$, $0 \le y \le b$ **(Fig. 9.17)**:

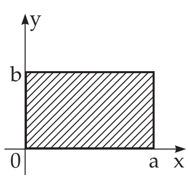


Figure 9.17

Find a function $u(x, y)$ satisfying the elliptic type Laplace differential equation

$$\Delta u = 0 \tag{9.95a}$$

and the boundary conditions

$$u(0, y) = \varphi_1(y), \quad u(a, y) = \varphi_2(y),$$
$$u(x, 0) = \psi_1(x), \quad u(x, b) = \psi_2(x). \tag{9.95b}$$

First there is to determine a particular solution for the boundary conditions $\varphi_1(y) = \varphi_2(y) = 0$. Substituting the product form

$$u = X(x)Y(y) \tag{9.95c}$$

into (9.95a) gives the separated differential equations

$$\frac{X''}{X} = -\frac{Y''}{Y} = -\lambda^2 \tag{9.95d}$$

with the eigenvalue λ analogously to examples **A** (p. 579) through **C** (p. 581). Since $X(0) = X(a) = 0$,

$$X = C\sin\lambda x, \quad \lambda = \frac{n\pi}{a} = \lambda_n \quad (n = 1, 2, \ldots). \tag{9.95e}$$

In the second step the general solution of the differential equation is obtained:

$$Y'' - \frac{n^2\pi^2}{a^2}Y = 0 \quad (9.95f) \qquad \text{in the form} \quad Y = a_n \sinh\frac{n\pi}{a}(b-y) + b_n \sinh\frac{n\pi}{a}y\,. \tag{9.95g}$$

From these equations one gets a particular solution of (9.95a) satisfying the boundary conditions $u(0, y) = u(a, y) = 0$, which has the form

$$u_n = \left[a_n \sinh\frac{n\pi}{a}(b-y) + b_n \sinh\frac{n\pi}{a}y\right]\sin\frac{n\pi}{a}x\,. \tag{9.95h}$$

In the third step one considers the general solution as a series

$$u = \sum_{n=1}^{\infty} u_n, \tag{9.95i}$$

so from the boundary conditions for $y = 0$ and $y = b$

$$u = \sum_{n=1}^{\infty}\left(a_n \sinh\frac{n\pi}{a}(b-y) + b_n \sinh\frac{n\pi}{a}y\right)\sin\frac{n\pi}{a}x \tag{9.95j}$$

follows with the coefficients

$$a_n = \frac{2}{a\sinh\frac{n\pi b}{a}}\int_0^a \psi_1(x)\sin\frac{n\pi}{a}x\,dx, \quad b_n = \frac{2}{a\sinh\frac{n\pi b}{a}}\int_0^a \psi_2(x)\sin\frac{n\pi}{a}x\,dx. \tag{9.95k}$$

The problem with the boundary conditions $\psi_1(x) = \psi_2(x) = 0$ can be solved in a similar manner, and taking the series (9.95j) one gets the general solution of (9.95a) and (9.95b).

■ **E: Heat Conduction Equation** Heat conduction in a homogeneous bar with one end at infinity and the other end kept at a constant temperature is described by the linear second-order partial differential equation of parabolic type

$$\frac{\partial u}{\partial t} = a^2\frac{\partial^2 u}{\partial x^2}, \tag{9.96a}$$

which satisfies the initial and boundary conditions

$$u|_{t=0} = f(x), \quad u|_{x=0} = 0 \tag{9.96b}$$

in the domain $0 \le x < +\infty$, $t \ge 0$. It is also to be supposed that the temperature tends to zero at infinity. Substituting

$$u = X(x)T(t) \tag{9.96c}$$

into (9.96a) one obtains the ordinary differential equations

$$\frac{T'}{a^2T} = \frac{X''}{X} = -\lambda^2, \tag{9.96d}$$

whose parameter λ is introduced analogously to the previous examples **A** (p. 579) through **D** (p. 582). One gets

$$T(t) = C_\lambda e^{-\lambda^2 a^2 t} \tag{9.96e}$$

as a solution for $T(t)$. Using the boundary condition $X(0) = 0$, gives

$$X(x) = C \sin \lambda x \quad (9.96f) \qquad \text{and so} \quad u_\lambda = C_\lambda e^{-\lambda^2 a^2 t} \sin \lambda x, \quad (9.96g)$$

where λ is an arbitrary real number. The solution can be obtained in the form

$$u(x,t) = \int_0^\infty C(\lambda) e^{-\lambda^2 a^2 t} \sin \lambda x \, d\lambda. \quad (9.96h)$$

From the initial condition $u|_{t=0} = f(x)$ follows (9.96i) with (9.96j) for the constant (see 7.4.1.1, **1.**, p. 474).

$$f(x) = \int_0^\infty C(\lambda) \sin \lambda x \, d\lambda\,, \quad (9.96i) \qquad C(\lambda) = \frac{2}{\pi} \int_0^\infty f(s) \sin \lambda s \, ds\,. \quad (9.96j)$$

Combining (9.96j) and (9.96h) gives

$$u(x,t) = \frac{2}{\pi} \int_0^\infty f(s) \left(\int_0^\infty e^{-\lambda^2 a^2 t} \sin \lambda s \sin \lambda x \, d\lambda \right) ds \quad (9.96k)$$

or after replacing the product of the two sines with one half of the difference of two cosines ((2.122), p. 83) and using formula (21.27), in **Table 21.8.2**, p. 1100, it follows that

$$u(x,t) = \int_0^\infty f(s) \frac{1}{2a\sqrt{\pi t}} \left[\exp\left(-\frac{(x-s)^2}{4a^2 t} \right) - \exp\left(-\frac{(x+s)^2}{4a^2 t} \right) \right] ds. \quad (9.96l)$$

2. Riemann Method for Solving Cauchy's Problem for the Hyperbolic Differential Equation

$$\frac{\partial^2 u}{\partial x \partial y} + a \frac{\partial u}{\partial x} + b \frac{\partial u}{\partial y} + cu = F \quad (9.97a)$$

1. Riemann Function is a function $v(x, y; \xi, \eta)$, where ξ and η are considered as parameters, satisfying the homogeneous equation

$$\frac{\partial^2 v}{\partial x \partial y} - \frac{\partial (av)}{\partial x} - \frac{\partial (bv)}{\partial y} + cv = 0 \quad (9.97b)$$

which is the adjoint of (9.97a) and the conditions

$$v(x, \eta; \xi, \eta) = \exp\left(\int_\xi^x b(s, \eta)\, ds \right), \quad v(\xi, y; \xi, \eta) = \exp\left(\int_\eta^y a(\xi, s)\, ds \right). \quad (9.97c)$$

In general, linear second-order differential equations and their adjoint differential equations have the form

$$\sum_{i,k} a_{ik} \frac{\partial^2 u}{\partial x_i \partial x_k} + \sum_i b_i \frac{\partial u}{\partial x_i} + cu = f \quad (9.97d) \qquad \text{and} \qquad \sum_{i,k} \frac{\partial^2 (a_{ik} v)}{\partial x_i \partial x_k} - \sum_i \frac{\partial (b_i v)}{\partial x_i} + cv = 0. \quad (9.97e)$$

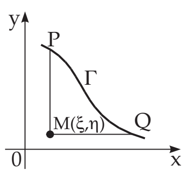


Figure 9.18

2. Riemann Formula is the integral formula which is used to determine function $u(\xi, \eta)$ satisfying the given differential equation (9.97a) and taking the previously given values along the previously given curve Γ **(Fig. 9.18)** together with its derivative in the direction of the curve normal (see 3.6.1.2, **2.**, p. 244):

$$u(\xi,\eta) = \frac{1}{2}(uv)_P + \frac{1}{2}(uv)_Q - \int_{\widehat{QP}} \left[buv + \frac{1}{2}\left(v\frac{\partial u}{\partial x} - u \frac{\partial v}{\partial x} \right) \right] dx$$

$$-\left[auv + \frac{1}{2}\left(v\frac{\partial u}{\partial y} - u\frac{\partial v}{\partial y}\right)\right] dy + \iint_{PMQ} Fv\,dx\,dy. \tag{9.97f}$$

The smooth curve Γ **(Fig. 9.18)** must not have tangents parallel to the coordinate axes, i.e., the curve must not be tangent to the characteristics. The line integral in this formula can be calculated, since the values of both partial derivatives can be determined from the function values and from its derivatives in a non-tangential direction along the curve arc.

In the Cauchy problem, the values of the partial derivatives of the unknown function, e.g., $\frac{\partial u}{\partial y}$ are often given instead of the normal derivative along the curve. Then another form of the Riemann formula is used:

$$u(\xi,\eta) = (uv)_P - \int_{\overset{\frown}{QP}} \left(buv - u\frac{\partial v}{\partial x}\right) dx - \left(auv + v\frac{\partial u}{\partial y}\right) dy + \iint_{PMQ} Fv\,dx\,dy. \tag{9.97g}$$

■ **Telegraph Equation** (Telegrapher's Equation) is a linear second-order partial differential equation of hyperbolic type

$$a\frac{\partial^2 u}{\partial t^2} + 2b\frac{\partial u}{\partial t} + cu = \frac{\partial^2 u}{\partial x^2} \tag{9.98a}$$

where $a > 0$, b, and c are constants. The equation describes the current flow in wires. It is a generalization of the differential equation of a vibrating string.
Replacing the unknown function $u(x,t)$ by $u = z\exp(-(b/a)t)$, so (9.98a) is reduced to the form

$$\frac{\partial^2 z}{\partial t^2} = m^2\frac{\partial^2 z}{\partial x^2} + n^2 z \quad \left(m^2 = \frac{1}{a},\ n^2 = \frac{b^2 - ac}{a^2}\right). \tag{9.98b}$$

Replacing the independent variables by

$$\xi = \frac{n}{m}(mt + x), \quad \eta = \frac{n}{m}(mt - x) \tag{9.98c}$$

finally one gets the canonical form

$$\frac{\partial^2 z}{\partial\xi\partial\eta} - \frac{z}{4} = 0 \tag{9.98d}$$

of a hyperbolic type linear partial differential equation (see 9.2.2.1, **1.**, p. 577). The Riemann function $v(\xi,\eta;\xi_0,\eta_0)$ should satisfy this equation with unit value at $\xi = \xi_0$ and $\eta = \eta_0$. Choosing the form

$$w = (\xi - \xi_0)(\eta - \eta_0) \tag{9.98e}$$

for w in $v = f(w)$, then $f(w)$ is a solution of the differential equation

$$w\frac{d^2 f}{dw^2} + \frac{df}{dw} - \frac{1}{4}f = 0 \tag{9.98f}$$

with initial condition $f(0) = 1$. The substitution $w = \alpha^2$ reduces this differential equation to Bessel's differential equation of order zero (see 9.1.2.6, **2.**, p. 562)

$$\frac{d^2 f}{d\alpha^2} + \frac{1}{\alpha}\frac{df}{d\alpha} - f = 0, \tag{9.98g}$$

hence the solution is

$$v = I_0\left[\sqrt{(\xi - \xi_0)(\eta - \eta_0)}\right]. \tag{9.98h}$$

A solution of the original differential equation (9.98a) satisfying the boundary conditions

$$z\Big|_{t=0} = f(x), \quad \frac{\partial z}{\partial t}\Big|_{t=0} = g(x) \tag{9.98i}$$

can be obtained substituting the found value of v into the Riemann formula and then returning to the original variables:

$$z(x,t) = \frac{1}{2}[f(x-mt) + f(x+mt)]$$
$$+\frac{1}{2}\int\limits_{x-mt}^{x+mt}\left[g(s)\frac{I_0\left(\frac{n}{m}\sqrt{m^2t^2-(s-x)^2}\right)}{m} - f(s)\frac{ntI_1\left(\frac{n}{m}\sqrt{m^2t^2-(s-x)^2}\right)}{\sqrt{m^2t^2-(s-x)^2}}\right]ds. \tag{9.98j}$$

3. Green's Method of Solving the Boundary Value Problem for Elliptic Differential Equations with Two Independent Variables

This method is very similar to the Riemann method of solving the Cauchy problem for hyperbolic differential equations.
If one wants to find a function $u(x,y)$ satisfying the elliptic type of linear second-order partial differential equation

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + a\frac{\partial u}{\partial x} + b\frac{\partial u}{\partial y} + c\,u = f \tag{9.99a}$$

in a given domain and taking the prescribed values on its boundary, first the *Green function* $G(x,y,\xi,\eta)$ has to be determined for this domain, where ξ and η are regarded as parameters. The Green function must satisfy the following conditions:

1. The function $G(x,y;\xi,\eta)$ satisfies the homogeneous adjoint differential equation

$$\frac{\partial^2 G}{\partial x^2} + \frac{\partial^2 G}{\partial y^2} - \frac{\partial(a\,G)}{\partial x} - \frac{\partial(b\,G)}{\partial y} + c\,G = 0 \tag{9.99b}$$

everywhere in the given domain except at the point $x=\xi$, $y=\eta$.

2. The function $G(x,y;\xi,\eta)$ has the form

$$U\ln\frac{1}{r} + V \tag{9.99c}$$ with $$r = \sqrt{(x-\xi)^2+(y-\eta)^2}\,, \tag{9.99d}$$

where U has unit value at the point $x=\xi$, $y=\eta$ and U and V are continuous functions in the entire domain together with their second derivatives.

3. The function $G(x,y;\xi,\eta)$ is equal to zero on the boundary of the given domain.

The second step is to give the solution of the boundary value problem with the Green function by the formula

$$u(\xi,\eta) = \frac{1}{2\pi}\int\limits_S u(x,y)\frac{\partial}{\partial n}G(x,y;\xi,\eta)\,ds - \frac{1}{2\pi}\iint\limits_D f(x,y)G(x,y;\xi,\eta)\,dx\,dy, \tag{9.99e}$$

where D is the considered domain, S is its boundary on which the function is assumed to be known and $\frac{\partial}{\partial n}$ denotes the normal derivative directed toward the interior of D.

Condition **3** depends on the formulation of the problem. For instance, if instead of the function values the values of the derivative of the unknown function are given in the direction normal to the boundary of the domain, then in **3** the condition

$$\frac{\partial G}{\partial n} - (a\cos\alpha + b\cos\beta)G = 0 \tag{9.99f}$$

holds on the boundary. α and β denote here the angles between the interior normal to the boundary of the domain and the coordinate axes. In this case, the solution is given by the formula

$$u(\xi,\eta) = -\frac{1}{2\pi}\int_S \frac{\partial u}{\partial n} G\,ds - \frac{1}{2\pi}\iint_D fG\,dx\,dy. \tag{9.99g}$$

4. Green's Method for the Solution of Boundary Value Problems with Three Independent Variables

The solution of the differential equation

$$\Delta u + a\frac{\partial u}{\partial x} + b\frac{\partial u}{\partial y} + c\frac{\partial u}{\partial z} + e\,u = f \tag{9.100a}$$

should take the given values on the boundary of the considered domain. As the first step, one constructs again the Green function, but now it depends on three parameters ξ, η, and ζ. The adjoint differential equation satisfied by the Green function has the form

$$\Delta G - \frac{\partial(a\,G)}{\partial x} - \frac{\partial(b\,G)}{\partial y} - \frac{\partial(c\,G)}{\partial z} + e\,G = 0. \tag{9.100b}$$

As in condition **2**, the function $G(x,y,z;\xi,\eta,\zeta)$ has the form

$$U\frac{1}{r} + V \tag{9.100c}$$

$$\text{with} \quad r = \sqrt{(x-\xi)^2 + (y-\eta)^2 + (z-\zeta)^2}. \tag{9.100d}$$

The solution of the problem is:

$$u(\xi,\eta,\zeta) = \frac{1}{4\pi}\iint_S u\frac{\partial G}{\partial n}\,ds - \frac{1}{4\pi}\iiint_D fG\,dx\,dy\,dz. \tag{9.100e}$$

Both methods, Riemann's and Green's, have the common idea first to determine a special solution of the differential equation, which can then be used to obtain a solution with arbitrary boundary conditions. An essential difference between the Riemann and the Green function is that the first one depends only on the form of the left-hand side of the differential equation, while the second one depends also on the considered domain. Finding the Green function is, in practice, an extremely difficult problem, even if it is known to exist; therefore, Green's method is used mostly in theoretical research.

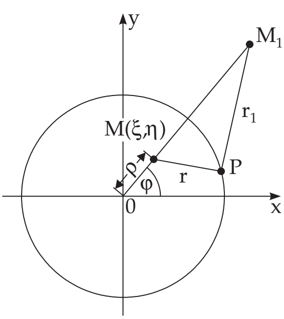


Figure 9.19

■ **A:** Construction of the Green function for the Dirichlet problem of the Laplace differential equation (see 13.5.1, p. 729)

$$\Delta u = 0 \tag{9.101a}$$

for the case, when the considered domain is a circle **(Fig. 9.19)**. The Green function is

$$G(x,y;\xi,\eta) = \ln\frac{1}{r} + \ln\frac{r_1\rho}{R}, \tag{9.101b}$$

where $r = \overline{MP}$, $\rho = \overline{OM}$, $r_1 = \overline{M_1P}$ and R is the radius of the considered circle **(Fig. 9.19)**. The points M and M_1 are symmetric with respect to the circle, i.e., both points are on the same ray starting from the center and

$$\overline{OM}\cdot\overline{OM_1} = R^2. \tag{9.101c}$$

The formula (9.99e) for a solution of Dirichlet's problem, after substituting the normal derivative of the Green function and after certain calculations, yields the so-called Poisson integral

$$u(\xi,\eta) = \frac{1}{2\pi}\int_0^{2\pi} \frac{R^2-\rho^2}{R^2+\rho^2-2R\rho\cos(\psi-\varphi)}u(\varphi)\,d\varphi. \tag{9.101d}$$

The notation is the same as above. The known values of u are given on the boundary of the circle by $u(\varphi)$. For the coordinates of the point $M(\xi, \eta)$ follows $\xi = \rho\cos\psi$, $\eta = \rho\sin\psi$.

■ **B:** Construction of the Green function for the Dirichlet problem of the Laplace differential equation (see 13.5.1, p. 729)

$$\Delta u = 0, \tag{9.102a}$$

for the case when the considered domain is a sphere with radius R. The Green function now has the form

$$G(x, y, z; \xi, \eta, \zeta) = \frac{1}{r} - \frac{R}{r_1 \rho}, \tag{9.102b}$$

with $\rho = \sqrt{\xi^2 + \eta^2 + \zeta^2}$ as the distance of the point (ξ, η, ζ) from the center, r as the distance between the points (x, y, z) and (ξ, η, ζ), and r_1 as the distance of the point (x, y, z) from the symmetric point of (ξ, η, ζ) according to (9.101c), i.e., from the point $\left(\frac{R\xi}{\rho}, \frac{R\eta}{\rho}, \frac{R\zeta}{\rho}\right)$. In this case, the Poisson integral has the form (with the same notation as in ■ **A** (p. 587))

$$u(\xi, \eta, \zeta) = \frac{1}{4\pi} \iint_S \frac{R^2 - \rho^2}{Rr^3} u \, ds. \tag{9.102c}$$

5. Operational Method

Operational methods can be used not only to solve ordinary differential equations but also for partial differential equations (see 15.1.6, p. 769). They are based on transition from the unknown function to its transform (see 15.1, p. 767). In this process the unknown function is regarded as a function of only one variable and the transformation is performed with respect to this variable. The remaining variables are considered as parameters. The differential equation to determine the transform of the unknown function contains one less independent variable than the original equation. In particular, if the original equation is a partial differential equation of two independent variables, then one obtains an ordinary differential equation for the transform. If the transform of the unknown function can be found from the obtained equation, then the original function is obtained either from the formula for the inverse function or from the table of transforms.

6. Approximation Methods

In order to solve practical problems with partial differential equations, different approximation methods are used. They can be divided into analytical and numerical methods.

1. Analytical Methods make possible the determination of approximate analytical expressions for the unknown function.

2. Numerical Methods result in approximate values of the unknown function for certain values of the independent variables. Here the following methods (see 19.5, p. 976) are used:

a) Finite Difference Method, or *Lattice-Point Method*: The derivatives are replaced by divided differences, so the differential equation including the initial and boundary conditions becomes an algebraic equation system. A linear differential equation with linear initial and boundary conditions becomes a linear equation system.

b) Finite Element Method, or briefly **FEM** (see 19.5.3, p. 978), for boundary value problems: Here a variational problem is assigned to the boundary value problem. The approximation of the unknown function is performed by a spline approach, whose coefficients should be chosen to get the best possible solution. The domain of the boundary value problem is decomposed into regular sub-domains. The coefficients are determined by solving an extreme value problem.

c) Integral Equation Method (along a Closed Curve) for special boundary problems: The boundary value problem is formulated as an equivalent integral equation problem along the boundary of the domain of the boundary value problem. To do this, one applies the theorems of vector analysis (see 13.3.3, p. 724, and followings), e.g., Green formulas. The remaining integrals along the closed curve are to be determined numerically by a suitable quadrature formula.

3. Physical Solutions of differential equations can be given by experimental methods. This is based on the fact that various physical phenomena can be described by the same differential equation. To solve a given equation, first a model is constructed by which the given problem is simulated and the values of the unknown function are obtained directly from this model. Since such models are often known and can be constructed by varying the parameters in a wide range, the differential equation can also be applied in a wide domain of the variables.

9.2.3 Some further Partial Differential Equations from Natural Sciences and Engineering

9.2.3.1 Formulation of the Problem and the Boundary Conditions

1. Problem Formulation

The modeling and the mathematical treatment of different physical phenomena in classical theoretical physics, especially in modeling media considered structureless or continuously changing, such as gases, fluids, solids, the fields of classical physics, leads to the introduction of partial differential equations. Examples are the wave (see 9.2.3.2, p. 590) and the heat equations (see 9.2.3.3, p. 591). Many problems in non-classical theoretical physics are also governed by partial differential equations. An important area is quantum mechanics, which is based on the recognition that media and fields are discontinuous. The most famous relation is the Schroedinger equation. Linear second-order partial differential equations occur most frequently and they have special importance in today's natural sciences.

2. Initial and Boundary Conditions

The solution of the problems of physics, engineering, and the natural sciences must usually fulfill two basic requirements:

1. The solution must satisfy not only the differential equation, but also certain initial and/or boundary conditions. There are problems with only initial condition or only with boundary conditions or with both. All the conditions together must determine the unique solution of the differential equation.

2. The solution must be stable with respect to small changes in the initial and boundary conditions, i.e., its change should be arbitrarily small if the *perturbations* of these conditions are small enough. Then a *correct problem formulation* is given.
One can assume that the mathematical model of the given problem to describe the real situation is adequate only in cases when these conditions are fulfilled.
For instance, the Cauchy problem (see 9.2.1.1, **5.**, p. 572) is correctly defined with a differential equation of hyperbolic type for investigating vibration processes in continuous media. This means that the values of the required function, and the values of its derivatives in a non-tangential (mostly in a normal) direction are given on an initial manifold, i.e., on a curve or on a surface.

In the case of differential equations of elliptic type, which occur in investigations of steady state and equilibrium problems in continuous media, the formulation of the boundary value problem is correct. If the considered domain is unbounded, then the unknown function must satisfy certain given properties with unlimited increase of the independent variables.

3. Inhomogeneous Conditions and Inhomogeneous Differential Equations

The solution of homogeneous or inhomogeneous linear partial differential equations with inhomogeneous initial or boundary conditions can be reduced to the solution of an equation which differs from the original one only by a free term not containing the unknown function, and which has homogeneous conditions. It is sufficient to replace the original function by its difference from an arbitrary twice differentiable function satisfying the given inhomogeneous conditions.

In general, one uses the fact that the solution of a linear inhomogeneous partial differential equation with given inhomogeneous initial or boundary conditions is the sum of the solutions of the same differential equation with zero conditions and the solution of the corresponding homogeneous differential equation with the given conditions.

To reduce the solution of the linear inhomogeneous partial differential equation

$$\frac{\partial^2 u}{\partial t^2} - L[u] = g(x,t) \tag{9.103a}$$

with homogeneous initial conditions

$$u\Big|_{t=0} = 0, \quad \frac{\partial u}{\partial t}\Big|_{t=0} = 0 \tag{9.103b}$$

to the solution of the Cauchy problem for the corresponding homogeneous differential equation, one substitutes

$$u = \int_0^t \varphi(x,t;\tau)\,d\tau. \tag{9.103c}$$

Here $\varphi(x,t;\tau)$ is the solution of the differential equation

$$\frac{\partial^2 u}{\partial t^2} - L[u] = 0, \tag{9.103d}$$

which satisfies the boundary conditions

$$u\Big|_{t=\tau} = 0, \quad \frac{\partial u}{\partial t}\Big|_{t=\tau} = g(x,\tau). \tag{9.103e}$$

In this equation, x represents symbolically all the n variables $x_1, x_2, \ldots, x_n$ of the n-dimensional problem. $L[u]$ denotes a linear differential expression, which may contain the derivative $\frac{\partial u}{\partial t}$, but not higher-order derivatives with respect to t.

9.2.3.2 Wave Equation

The extension of oscillations in a homogeneous media is described by the *wave equation*

$$\frac{\partial^2 u}{\partial t^2} - a^2\Delta u = Q(x,t), \tag{9.104a}$$

whose right-hand side $Q(x,t)$ vanishes when there is no perturbation. The symbol x represents the n variables $x_1, \ldots, x_n$ of the n-dimensional problem. The Laplace operator Δ (see also 13.2.6.5, 716,) is defined in the following way:

$$\Delta u = \frac{\partial^2 u}{\partial {x_1}^2} + \frac{\partial^2 u}{\partial {x_2}^2} + \cdots + \frac{\partial^2 u}{\partial {x_n}^2}. \tag{9.104b}$$

The solution of the wave equation is the *wave function* u. The differential equation (9.104a) is of hyperbolic type.

1. Homogeneous Problem

The solution of the homogeneous problem with $Q(x,t) = 0$ and with the initial conditions

$$u\Big|_{t=0} = \varphi(x), \quad \frac{\partial u}{\partial t}\Big|_{t=0} = \psi(x) \tag{9.105}$$

is given for the cases $n = 1, 2, 3$ by the following integrals.

Case $n = 3$ (Kirchhoff Formula):

$$u(x_1,x_2,x_3,t) = \frac{1}{4\pi a^2}\left[\iint\limits_{(S_{\rm at})} \frac{\psi(\alpha_1,\alpha_2,\alpha_3)}{t}\,d\sigma + \frac{\partial}{\partial t}\iint\limits_{(S_{\rm at})} \frac{\varphi(\alpha_1,\alpha_2,\alpha_3)}{t}\,d\sigma\right], \tag{9.106a}$$

where the integration is performed over the spherical surface S_{at} given by the equation $(\alpha_1 - x_1)^2 + (\alpha_2 - x_2)^2 + (\alpha_3 - x_3)^2 = a^2t^2$.

Case $n = 2$ (Poisson Formula):

$$u(x_1, x_2, t) = \frac{1}{2\pi a}\Big[\iint\limits_{(C_{\text{at}})} \frac{\psi(\alpha_1, \alpha_2)\, d\alpha_1 d\alpha_2}{\sqrt{a^2t^2 - (\alpha_1 - x_1)^2 - (\alpha_2 - x_2)^2}} + \frac{\partial}{\partial t} \iint\limits_{(C_{\text{at}})} \frac{\varphi(\alpha_1, \alpha_2)\, d\alpha_1 d\alpha_2}{\sqrt{a^2t^2 - (\alpha_1 - x_1)^2 - (\alpha_2 - x_2)^2}}\Big], \tag{9.106b}$$

where the integration is performed along the circle C_{at} given by the equation $(\alpha_1 - x_1)^2 + (\alpha_2 - x_2)^2 \leq a^2t^2$.

Case $n = 1$ (d'Alembert formula):

$$u(x_1, t) = \frac{\varphi(x_1 + at) + \varphi(x_1 - at)}{2} + \frac{1}{2a} \int\limits_{x_1 - at}^{x_1 + at} \psi(\alpha)\, d\alpha. \tag{9.106c}$$

2. Inhomogeneous Problem

In the case, when $Q(x, t) \neq 0$, one has to add to the right-hand sides of (9.106a,b,c) the correcting terms:

Case $n = 3$ (Retarded Potential): For a domain K given by $r \leq at$ with $r = \sqrt{(\xi_1 - x_1)^2 + (\xi_2 - x_2)^2 + (\xi_3 - x_3)^2}$, the correction term is

$$\frac{1}{4\pi a^2} \iiint\limits_{(K)} \frac{Q\left(\xi_1, \xi_2, \xi_3, t - \frac{r}{a}\right)}{r}\, d\xi_1 d\xi_2 d\xi_3\,. \tag{9.107a}$$

Case $n = 2$: $$\frac{1}{2\pi a} \iiint\limits_{(K)} \frac{Q(\xi_1, \xi_2, \tau)\, d\xi_1 d\xi_2 d\tau}{\sqrt{a^2(t - \tau)^2 - (\xi_1 - x_1)^2 - (\xi_2 - x_2)^2}}\,, \tag{9.107b}$$

where K is a domain of ξ_1, ξ_2, τ space defined by the inequalities $0 \leq \tau \leq t$, $(\xi_1 - x_1)^2 + (\xi_2 - x_2)^2 \leq a^2(t - \tau)^2$.

Case $n = 1$: $$\frac{1}{2a} \iint\limits_{(T)} Q(\xi, \tau)\, d\xi d\tau, \tag{9.107c}$$

where T is the triangle $0 \leq \tau \leq t$, $|\xi - x_1| \leq a|t - \tau|$. a denotes the wave velocity of the perturbation.

9.2.3.3 Heat Conduction and Diffusion Equation for Homogeneous Media

1. Three-Dimensional Heat Conduction Equation

The propagation of heat in a homogeneous medium is described by a linear second-order partial differential equation of parabolic type

$$\frac{\partial u}{\partial t} - a^2 \Delta u = Q(x, t), \tag{9.108a}$$

where Δ is the three-dimensional Laplace operator defined in three directions of propagation x_1, x_2, x_3, determined by the position vector $\vec{\mathbf{r}}$. If the heat flow has neither source nor sink, the right-hand side vanishes since $Q(x, t) = 0$.

The Cauchy problem can be posed in the following way: It is to determine a bounded solution $u(x, t)$ for

$t > 0$, where $u|_{t=0} = f(x)$. The requirement of boundedness guarantees the uniqueness of the solution. For the homogeneous differential equation with $Q(x,t) = 0$, one gets the *wave function*

$$u(x_1, x_2, x_3, t) = \frac{1}{(2a\sqrt{\pi t})^n} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(\alpha_1, \alpha_2, \alpha_3) \cdot \exp\left(-\frac{(x_1-\alpha_1)^2 + (x_2-\alpha_2)^2 + (x_3-\alpha_3)^2}{4a^2 t}\right) d\alpha_1 d\alpha_2 d\alpha_3. \tag{9.108b}$$

In the case of an inhomogeneous differential equation with $Q(x,t) \neq 0$, one has to add to the right-hand side of (9.108b) the following expression:

$$\int_0^t \left[\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \frac{Q(\alpha_1, \alpha_2, \alpha_3)}{[2a\sqrt{\pi(t-\tau)}]^n} \cdot \exp\left(-\frac{(x_1-\alpha_1)^2 + (x_2-\alpha_2)^2 + (x_3-\alpha_3)^2}{4a^2(t-\tau)}\right) d\alpha_1 d\alpha_2 d\alpha_3 \right] d\tau. \tag{9.108c}$$

The problem of determining $u(x,t)$ for $t < 0$, if the values $u(x,0)$ are given, cannot be solved in this way, since the Cauchy problem is not correctly formulated in this case.
Since the temperature difference is proportional to the heat, one often introduces $u = T(\vec{\mathbf{r}}, t)$ (temperature field) and $a^2 = D_W$ (heat diffusion constant or thermal conductivity) to get

$$\frac{\partial T}{\partial t} - D_W \Delta T = Q_W(\vec{\mathbf{r}}, t). \tag{9.108d}$$

2. Three-Dimensional Diffusion Equation

In analogy to the heat equation, the propagation of a concentration C in a homogeneous medium is described by the same linear partial differential equation (9.108a) and (9.108d), where D_W is replaced by the three-dimensional *diffusion coefficient* D_C. The *diffusion equation* is:

$$\frac{\partial C}{\partial t} - D_C \Delta C = Q_C(\vec{\mathbf{r}}, t). \tag{9.109}$$

One gets the solutions by changing the symbols in the wave equations (9.108b) and (9.108c).

9.2.3.4 Potential Equation

The linear second-order partial differential equation

$$\Delta u = -4\pi\varrho \tag{9.110a}$$

is called the *potential equation* or *Poisson differential equation* (see 13.5.2, p. 729), which makes the determination of the potential $u(x)$ of a scalar field determined by a scalar point function $\varrho(x)$ possible, where x has the coordinates x_1, x_2, x_3 and Δ is the Laplace operator. The solution, the potential $u_M(x_1, x_2, x_3)$ at the point M, is discussed in 13.5.2, p. 729.
One gets the *Laplace differential equation* (see 13.5.1, p. 729) for the homogeneous differential equation with $\varrho \equiv 0$:

$$\Delta u = 0. \tag{9.110b}$$

The differential equations (9.110a) and (9.110b) are of elliptic type.

9.2.4 Schroedinger's Equation

9.2.4.1 Notion of the Schroedinger Equation

1. Determination and Dependencies

The solutions of the Schroedinger equation, the *wave functions* ψ, describe the properties of a quantum mechanical system, i.e., the properties of the states of a particle. The Schroedinger equation is a second-

order partial differential equation with the second-order derivatives of the wave function with respect to the space coordinates and first-order with respect to the time coordinate:

$$\mathrm{i}\,\hbar\frac{\partial\psi}{\partial t} = -\frac{\hbar^2}{2m}\Delta\psi + U(x_1,x_2,x_3,t)\,\psi = \hat{H}\,\psi \tag{9.111a}$$

$$\hat{H} \equiv \frac{\hat{p}^2}{2m} + U(\vec{\mathbf{r}},t)\,, \quad \hat{p} \equiv \frac{\hbar}{\mathrm{i}}\frac{\partial}{\partial\vec{\mathbf{r}}} \equiv \frac{\hbar}{\mathrm{i}}\nabla. \tag{9.111b}$$

Here, Δ is the Laplace operator, $\hbar = \dfrac{h}{2\pi}$ is the reduced Planck's constant, i is the imaginary unit and ∇ is the nabla operator. The relation between the impulse p of a free particle with mass m and wave length λ is $\lambda = h/p$.

2. Remarks:

a) In quantum mechanics, an operator is assigned to every measurable quantity. The operator occurring in (9.111a) and (9.111b) is called the *Hamilton operator* $\hat{H}$ ("*Hamiltonian*") . It has the same role as the Hamilton function of classical mechanical systems (see, e.g., the example on Two-Body Problem on p. 574). It represents the total energy of the system which is divided into kinetic and potential energy. The first term in $\hat{H}$ is the operator for the kinetic energy, the second one for the potential energy.

b) The imaginary unit appears explicitly in the Schroedinger equation. Consequently, the wave functions are complex functions. Both real functions occurring in $\psi^{(1)} + \mathrm{i}\psi^{(2)}$ are needed to calculate the observable quantities. The square $|\Psi|^2$ of the wave function, describing the probability dw of the particle being in an arbitrary volume element dV of the observed domain, must satisfy special further conditions.

c) Besides the potential of the *interaction*, every special solution depends also on the initial and boundary conditions of the given problem. In general, there is a linear second-order boundary value problem, whose solutions have physical meaning only for the eigenvalues. The squares of the absolute value of meaningful solutions are everywhere unique and regular, and tend to zero at infinity.

d) The micro-particles also have wave and particle properties based on the *wave–particle duality*, so the Schroedinger equation is a wave equation (see 9.2.3.2, p. 590) for the De Broglie matter waves.

e) The restriction to the non-relativistic case means that the velocity v of the particle is very small with respect to the velocity of light c ($v \ll c$).

The application of the Schroedinger equations is discussed in detail in the literature of theoretical physics (see, e.g., [9.15], [9.7],[9.10], [22.15]). In this chapter only some most important examples are demonstrated.

9.2.4.2 Time-Dependent Schroedinger Equation

The time-dependent Schroedinger equation (9.111a) describes the general non-relativistic case of a spin-less particle with mass m in a position-dependent and time-dependent potential field $U(x_1,x_2,x_3,t)$. The special conditions, which must be satisfied by the wave function, are:

a) The function ψ must be bounded and continuous.

b) The partial derivatives $\partial\psi/\partial x_1$, $\partial\psi/\partial x_2$, and $\partial\psi/\partial x_3$ must be continuous.

c) The function $|\psi|^2$ must be integrable, i.e.,

$$\iiint\limits_V |\psi(x_1,x_2,x_3,t)|^2\,dV < \infty. \tag{9.112a}$$

According to the normalization condition, the probability that the particle is in the considered domain must be equal to one. (9.112a) is sufficient to guarantee the condition, since multiplying ψ by an appropriate constant the value of the integral becomes one.

A solution of the time-dependent Schroedinger equation has the form

$$\psi(x_1,x_2,x_3,t) = \Psi(x_1,x_2,x_3)\exp\left(-\mathrm{i}\tfrac{E}{\hbar}t\right)\,. \tag{9.112b}$$

The state of the particle is described by a periodic function of time with angular frequency $\omega = E/\hbar$. If the energy of the particle has the fixed value $E = \text{const}$, then the probability dw of finding the particle in a space element dV is independent of time:

$$d\omega = |\psi|^2\, dV = \psi\psi^*\, dV. \tag{9.112c}$$

Then one speaks about a *stationary state* of the particle.

9.2.4.3 Time-Independent Schroedinger Equation

If the potential U does not depend on time, i.e., $U = U(x_1, x_2, x_3)$, then it is the time-independent Schroedinger equation and the wave function $\Psi(x_1, x_2, x_3)$ is sufficient to describe the state. Reducing it from the time-dependent Schroedinger equation (9.111a) with the solution (9.112b) gives

$$\Delta\Psi + \frac{2m}{\hbar^2}(E - U)\Psi = 0. \tag{9.113a}$$

In this non-relativistic case, the energy of the particle is

$$E = \frac{p^2}{2m} \quad \left(p = \frac{h}{\lambda}, \quad h = 2\pi\hbar\right). \tag{9.113b}$$

The wave functions Ψ satisfying this differential equation are the *eigenfunctions*; they exist only for certain *energy values* E, which are given for the considered problem of the special boundary conditions. The union of the eigenvalues forms the *energy spectrum* of the particle. If U is a potential of finite depth and it tends to zero at infinity, then the negative eigenvalues form a *discrete spectrum*.

If the considered domain is the entire space, then it can be required as a boundary condition that Ψ is quadratically integrable in the entire space in the Lebesgue sense (see 12.9.3.2, p. 696 and [8.5]). If the domain is finite, e.g., a sphere or a cylinder, then one can require, e.g., $\Psi = 0$ for the boundary as the first boundary condition problem.
This gives the *Helmholtz differential equation* in the special case of $U(x) = 0$:

$$\Delta\Psi + \lambda\Psi = 0 \tag{9.114a}$$

$$\text{with the eigenvalue} \quad \lambda = \frac{2mE}{\hbar^2}. \tag{9.114b}$$

$\Psi = 0$ is often required here as a boundary condition. (9.114a) represents the initial mathematical equation for acoustic oscillation in a finite domain.

9.2.4.4 Statistical Interpretation of the Wave Function

The quantum mechanics postulates that the complete description of a regarded single-particle system in the time t is to be performed by the complex *wave function* $\psi(\vec{\mathbf{r}}, t)$ as a *state function* and normalized solution of the Schroedinger equation. So, the wave function contains all possible experimental information, which can be got by measurements on this system. There exist no hidden sub-structures of the theory and no hidden parameters which could eliminate the *principal statistical character* of quantum mechanics, as it contains the connection of state function and ψ and measurement results.

1. Observable and Probability Amplitude

A physical expression (position, momentum, angular momentum, energy), which can be determined by a suitable measuring instrument, is called an *observable*. In quantum mechanics every observable A is represented by a linear, hermitian operator $\hat{A}$ with $\hat{A}^+ = \hat{A}$, which interacts on the wave function. At the same time the operator of the quantum mechanics takes over the structure of the classical expression.

■ For the operator $\hat{\vec{\mathbf{l}}}$ of the angular momentum, where $\hat{\vec{\mathbf{r}}}$ is the position operator $\hat{\vec{\mathbf{p}}}$ the momentum operator:

$$\hat{\vec{\mathbf{l}}} = (\hat{l}_x, \hat{l}_y, \hat{l}_z) = \hat{\vec{\mathbf{r}}} \times \hat{\vec{\mathbf{p}}} \quad \text{i.e.,} \qquad \hat{l}_x = \hat{y}\hat{p}_z - \hat{z}\hat{p}_y = \frac{\hbar}{i}\left(y\frac{\partial}{\partial z} - z\frac{\partial}{\partial y}\right), \tag{9.115a}$$

$$\hat{l}_y = \hat{z}\hat{p}_x - \hat{x}\hat{p}_z = \frac{\hbar}{i}\left(z\frac{\partial}{\partial x} - x\frac{\partial}{\partial z}\right), \quad \hat{l}_z = \hat{x}\hat{p}_y - \hat{y}\hat{p}_x = \frac{\hbar}{i}\left(x\frac{\partial}{\partial y} - y\frac{\partial}{\partial x}\right). \tag{9.115b}$$

In general, it is not possible to assign a certain numerical value to an observable by determining the wave function, but first only as the result of a measurement. The only possible measurement values A are the real eigenvalues a_i of $\hat{A}$; the associated eigenfunctions φ_i form a complete orthogonal system:

$$\hat{A}\varphi_i = a_i\,\varphi_i \quad (i,k = 1,2,\dots), \qquad \iiint\limits_V \varphi_i^*\varphi_k\,dV = \delta_{i,k}\,. \tag{9.116}$$

If the system is in an arbitrarily general state ψ, the result of a single experiment, i.e. the occurrence of a certain measure value a_i in a single measurement can not be predicted. If imaging to perform the measurement on $N \to \infty$ identical systems being in the same state ψ, then among the measurement results every possible result a_i can be found with a frequency N_i. The probability W_i to find the value a_i in a single measurement can be determined:

$$W_i = \lim_{N\to\infty}\frac{N_i}{N}, \qquad \sum_i N_i = N\,. \tag{9.117}$$

To determine this probability from the wave function ψ, one performs an expansion of ψ as a series of eigenfunctions φ_i:

$$\psi = \sum_i c_i\,\varphi_i\,, \qquad c_i = \iiint\limits_V \varphi_i^*\psi\,dV\,. \tag{9.118}$$

The coefficient of expansion c_i is the probability to find the system ψ in its characteristic state φ_i, i.e., to obtain the measuring value a_i. From the absolute square of c_i one gets the probability W_i for the measuring value a_i:

$$W_i = |c_i|^2\,, \quad \sum_i W_i = \iiint\limits_V \psi^*\psi\,dV = 1\,. \tag{9.119}$$

Because in every measurement it is sure to find one of the possible measuring values a_i, the sum of the probabilities W_i fulfills the condition of normalization for the wave function ψ.

If two states ψ_1, ψ_2 of a physical system are known, then from the linearity of the Schroedinger equation follows, that the superposition

$$\psi = \psi_1 + \psi_2 \tag{9.120}$$

also represents a possible physical state. This fundamental *superposition principle* of quantum mechanics is the reason why at the determination of probabilities with the state function ψ, e.g.,

$$|\psi|^2 = |\psi_1 + \psi_2|^2 = |\psi_1|^2 + |\psi_2|^2 + 2\mathrm{Re}(\psi_1\,\psi_2^*) \tag{9.121}$$

besides the single probabilities $|\psi_1|^2\,, |\psi_2|^2$ occurs an additional term with sign. This explains the surprising interference effects of quantum mechanics, e.g. (*wave-particle duality*).

2. Expectation Value and Uncertainty

The *quantum mechanical expectation value* $\overline{A}$ is defined as the mean value of the measurement results obtained from measurements with $N \to \infty$ identical systems:

$$\overline{A} = \lim_{N\to\infty}\frac{1}{N}\sum_i a_i\,N_i = \sum_i a_i\,W_i = \iiint\limits_V \psi^*\,\hat{A}\,\psi\,dV\,. \tag{9.122}$$

The expectation value usually is not identical to a possible measurement result.

■ The calculation of the expectation value $\overline{\vec{\mathbf{r}}} = (\overline{x}, \overline{y}, \overline{z})$ of a position measurement for a particle in the state $\psi(\vec{\mathbf{r}}, t)$, e.g.

$$\overline{x} = \iiint\limits_V x|\psi(\vec{\mathbf{r}}, t)|^2\,dV$$

shows, that the wave function $\psi(\vec{\mathbf{r}}, t)$ is to be interpreted as a probability amplitude. The absolute square $|\psi|^2$ then is a probability density. The expression

$$dW = |\psi(\vec{\mathbf{r}}, t)|^2\, dV\,, \qquad \int dW = \iiint_V |\psi(\vec{\mathbf{r}}, t)|^2\, dV = 1$$

is to be understood as a probability, to find the particle at the time t in the volume element dV in the position $\vec{\mathbf{r}}$ (*probability of the position*).

As a measure of the distribution of the measured results for an observable A, given for a general state in some measurements, can be defined by the help of the so-called *uncertainty* $\Delta\, A$ near the expectation value $\overline{A}$, which is to be introduced via the standard error:

$$(\Delta\, A)^2 = \lim_{N\to\infty} \frac{1}{N} \sum_i N_i\,(a_i - \overline{A})^2 = \sum_i W_i\,(a_i - \overline{A})^2\,. \qquad (9.123)$$

By the help of the wave function ψ one can determine the uncertainty $\Delta\, A$ of a observable as expectation value of the quadratic deviation from the mean value $\overline{A}$:

$$(\Delta\, A)^2 = \overline{(A - \overline{A})^2} = \overline{A^2} - \overline{A}^2 = \iiint_V \psi^*(\hat{A} - \overline{A})^2 \psi\, dV\,. \qquad (9.124)$$

If the system is in an eigenstate φ_i of $\hat{A}$, then all measurements give the same measuring value a_i:

$$\overline{A} = a_i, \qquad \Delta\, A = 0\,. \qquad (9.125)$$

A distribution near the expectation value $\overline{A}$ does not appear.

3. Uncertainty Relation

Considering two observable $A\,, B\,,$ whose operators commutate (see Lie brackets in 5.3.6.4, **2.**, p. 356),

$$\hat{C} = [\hat{A}\,, \hat{B}] = \hat{A}\hat{B} - \hat{B}\hat{A} = 0, \qquad (9.126)$$

then (and only then) exists a simultaneous system of eigenfunctions $\varphi_{i,\nu}$ $(i, \nu = 1, 2, \ldots)$

$$\hat{A}\varphi_{i,\nu} = a_i\, \varphi_{i,\nu} \qquad \hat{B}\varphi_{i,\nu} = b_\nu\, \varphi_{i,\nu}\,. \qquad (9.127)$$

In this case exist physical states, in which the expected values of both operators are eigenvalues, so that the uncertainties $\Delta\, A, \Delta\, B$ simultaneously disappear:

$$\overline{A} = a_i\,, \qquad \overline{B} = b_\nu\,, \qquad \Delta\, A = \Delta\, B = 0. \qquad (9.128)$$

Performing in this system a measurement of the observable A, which leads with the measured value a_i to the state $\varphi_{i,\nu}$, then a following measurement of B gives the measured value b_ν, without interfering the state generated in the first measurement (compatible observable, tolerance measurement).

For two observable A and B, which are represented by non-commutative operators there does not exist a simultaneous system of eigenfunctions. In this case it is impossible to find a physical state, for which the uncertainties $\Delta\, A, \Delta\, B$ can be simultaneously arbitrarily small. For the product of the uncertainties exists a lower bound, defined by the expectation value of the *commutator* $\hat{C}$:

$$\Delta\, A\, \Delta\, B \geq \left| \frac{1}{2i}\, \overline{[\hat{A}, \hat{B}]} \right|\,. \qquad (9.129)$$

This relation is called the *uncertainty relation*. The commutation relation (9.130) (see also 5.3.6.4, **2.**, p. 356) between the components, e.g., of the position and the momentum operator into the same direction

$$[\hat{p}_x\,, \hat{x}] = \frac{\hbar}{i} \qquad (9.130) \qquad\qquad \Delta\, p_x\, \Delta\, x \geq \frac{\hbar}{2}\,. \qquad (9.131)$$

leads to the Heisenberg uncertainty relation (9.131). In other words: there is a fundamental limitation

on how precisely both the position and the momentum of a particle can be simultaneously known.

9.2.4.5 Force-Free Motion of a Particle in a Block

1. Formulation of the Problem

A particle with a mass m is moving freely in a block with impenetrable walls of edge lengths a, b, c, therefore, it is in a potential box which is infinitely high in all three directions because of the impenetrability of the walls. That is, the probability of the presence of the particle, and also the wave function Ψ, vanishes outside the box. The Schroedinger equation and the boundary conditions for this problem are

$$\frac{\partial^2\Psi}{\partial x^2} + \frac{\partial^2\Psi}{\partial y^2} + \frac{\partial^2\Psi}{\partial z^2} + \frac{2m}{\hbar^2}E\Psi = 0, \quad (9.132a) \qquad \Psi = 0 \quad \text{for} \quad \begin{cases} x = 0, & x = a, \\ y = 0, & y = b, \\ z = 0, & z = c. \end{cases} \quad (9.132b)$$

2. Solution Approach

Separating the variables

$$\Psi(x, y, z) = \Psi_x(x)\,\Psi_y(y)\,\Psi_z(z) \quad (9.133a)$$

and substituting into (9.132a) gives

$$\frac{1}{\Psi_x}\frac{d^2\Psi_x}{dx^2} + \frac{1}{\Psi_y}\frac{d^2\Psi_y}{dy^2} + \frac{1}{\Psi_z}\frac{d^2\Psi_z}{dz^2} = -\frac{2m}{\hbar^2}E = -B. \quad (9.133b)$$

Every term on the left-hand side depends only on one independent variable. Their sum can be a constant $-B$ for arbitrary x, y, z only if every single term is a constant. In this case the partial differential equation is reduced to three ordinary differential equations:

$$\frac{d^2\Psi_x}{dx^2} = -{k_x}^2\Psi_x, \quad \frac{d^2\Psi_y}{dy^2} = -{k_y}^2\Psi_y, \quad \frac{d^2\Psi_z}{dz^2} = -{k_z}^2\Psi_z. \quad (9.133c)$$

The relation for the *separation constants* $-{k_x}^2$, $-{k_y}^2$, $-{k_z}^2$ is

$${k_x}^2 + {k_y}^2 + {k_z}^2 = B, \quad (9.133d) \qquad \text{consequently} \quad E = \frac{\hbar^2}{2m}({k_x}^2 + {k_y}^2 + {k_z}^2). \quad (9.133e)$$

3. Solutions

of the three equations (9.133c) are the functions

$$\Psi_x = A_x \sin k_x\, x, \quad \Psi_y = A_y \sin k_y\, y, \quad \Psi_z = A_z \sin k_z\, z \quad (9.134a)$$

with the constants A_x, A_y, A_z. With these functions Ψ satisfies the boundary conditions $\Psi = 0$ for $x = 0$, $y = 0$ and $z = 0$.

$$\sin k_x\, a = \sin k_y\, b = \sin k_z\, c = 0 \quad (9.134b)$$

must be valid to satisfy also the relation $\Psi = 0$ for $x = a$, $y = b$ and $z = c$, i.e., the relations

$$k_x = \frac{\pi n_x}{a}, \quad k_y = \frac{\pi n_y}{b}, \quad k_z = \frac{\pi n_z}{c} \quad (9.134c)$$

must be satisfied, where n_x, n_y, and n_z are integers.
One gets for the total energy

$$E_{n_x,n_y,n_z} = \frac{\hbar^2}{2m}\left[\left(\frac{n_x}{a}\right)^2 + \left(\frac{n_y}{b}\right)^2 + \left(\frac{n_z}{c}\right)^2\right] \quad (n_x, n_y, n_z = \pm 1, \pm 2, \ldots). \quad (9.134d)$$

It follows from this formula that the changes of energy of a particle by interchange with the neighborhood is not continuous, which is possible only in quantum systems. The numbers n_x, n_y, and n_z, belonging to the *eigenvalues* of the energy, are called the *quantum numbers*.

After calculating the product of constants $A_xA_yA_z$ from the *normalization condition*

$$(A_xA_yA_z)^2 \int_0^a\int_0^b\int_0^c \sin^2\frac{\pi n_x x}{a}\,\sin^2\frac{\pi n_y y}{b}\,\sin^2\frac{\pi n_z z}{c}\,dx\,dy\,dz = 1 \qquad (9.134e)$$

one gets the complete *eigenfunctions* of the states characterized by the three quantum numbers

$$\Psi_{n_x,n_y,n_z} = \sqrt{\frac{8}{abc}}\,\sin\frac{\pi n_x x}{a}\,\sin\frac{\pi n_y y}{b}\,\sin\frac{\pi n_z z}{c}\,. \qquad (9.134f)$$

The eigenfunctions vanish at the walls since one of the three sine functions is equal to zero. This is always the case outside the walls if the following relations are valid

$$x = \frac{a}{n_x}, \frac{2a}{n_x}, \ldots, \frac{(n_x-1)a}{n_x}, \quad y = \frac{b}{n_y}, \frac{2b}{n_y}, \ldots, \frac{(n_y-1)b}{n_y}, \quad z = \frac{c}{n_z}, \frac{2c}{n_z}, \ldots, \frac{(n_z-1)c}{n_z}. \qquad (9.134g)$$

So, there are $n_x - 1$ and $n_y - 1$ and $n_z - 1$ planes perpendicular to the x- or y- or z-axis, in which Ψ vanishes. These planes are called the *nodal planes*.

4. Special Case of a Cube, Degeneracy

In the special case of a cube with $a = b = c$, a particle can be in different states which are described by different linearly independent eigenfunctions and they have the same energy. This is the case when the sum ${n_x}^2 + {n_y}^2 + {n_z}^2$ has the same value in different states. They are called *degenerate states*, and if there are i states with the same energy, they are called *i-fold degeneracy*.
The quantum numbers n_x, n_y and n_z can run through all real numbers, except zero. This last case would mean that the wave function is identically zero, i.e., the particle does not exist at any place in the box. The particle energy must remain finite, even if the temperature reaches absolute zero. This *zero-point translational energy* for a block is

$$E_0 = \frac{\hbar^2}{2m}\left(\frac{1}{a^2} + \frac{1}{b^2} + \frac{1}{c^2}\right). \qquad (9.134h)$$

9.2.4.6 Particle Movement in a Symmetric Central Field (see 13.1.2.2, p. 702)

1. Formulation of the Problem

The considered particle moves in a central symmetric potential $V(r)$. This model reproduces the movement of an electron in the electrostatic field of a positively charged nucleus. Since this is a spherically symmetric problem, it is reasonable to use spherical coordinates **(Fig. 9.20)**. The following relations hold:

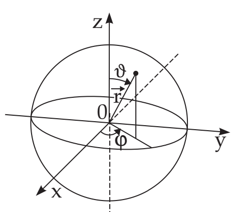


Figure 9.20

$$r = \sqrt{x^2 + y^2 + z^2}\,, \qquad x = r\sin\vartheta\cos\varphi,$$
$$\vartheta = \arccos\frac{z}{r}\,, \qquad y = r\sin\vartheta\sin\varphi, \qquad (9.135a)$$
$$\varphi = \arctan\frac{y}{x}\,, \qquad z = r\cos\vartheta,$$

where r is the absolute value of the radius vector, ϑ is the angle between the radius vector and the z-axis (polar angle) and φ is the angle between the projection of the radius vector onto the x, y plane and the x-axis (azimuthal angle). For the Laplace operator

$$\Delta\Psi = \frac{\partial^2\Psi}{\partial r^2} + \frac{2}{r}\frac{\partial\Psi}{\partial r} + \frac{1}{r^2}\frac{\partial^2\Psi}{\partial\vartheta^2} + \frac{\cos\vartheta}{r^2\sin\vartheta}\frac{\partial\Psi}{\partial\vartheta} + \frac{1}{r^2\sin^2\vartheta}\frac{\partial^2\Psi}{\partial\varphi^2}\,, \qquad (9.135b)$$

holds, so the time-independent Schroedinger equation is:

$$\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial\Psi}{\partial r}\right) + \frac{1}{r^2\sin\vartheta}\frac{\partial}{\partial\vartheta}\left(\sin\vartheta\frac{\partial\Psi}{\partial\vartheta}\right) + \frac{1}{r^2\sin^2\vartheta}\frac{\partial^2\Psi}{\partial\varphi^2} + \frac{2m}{\hbar^2}[E - V(r)]\Psi = 0. \qquad (9.135c)$$

2. Solution

Looking for a solution in the form

$$\Psi(r,\vartheta,\varphi) = R_l(r)Y_l^m(\vartheta,\varphi), \tag{9.136a}$$

where R_l is the radial wave function depending only on r, and $Y_l^m(\vartheta,\varphi)$ is the wave function depending on both angles. Substituting (9.136a) in (9.135c) gives

$$\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial R_l}{\partial r}\right)Y_l^m + \frac{2m}{\hbar^2}[E - V(r)]R_lY_l^m$$
$$= -\left\{\frac{1}{r^2\sin\vartheta}\frac{\partial}{\partial\vartheta}\left(\sin\vartheta\frac{\partial Y_l^m}{\partial\vartheta}\right)R_l + \frac{1}{r^2\sin^2\vartheta}\frac{\partial^2 Y_L^m}{\partial\varphi^2}R_l\right\}. \tag{9.136b}$$

Dividing by $R_lY_l^m$ and multiplying by r^2 gives

$$\frac{1}{R_l}\frac{d}{dr}\left(r^2\frac{dR_l}{dr}\right) + \frac{2mr^2}{\hbar^2}[E - V(r)] = -\frac{1}{Y_l^m}\left\{\frac{1}{\sin\vartheta}\frac{\partial}{\partial\vartheta}\left(\sin\vartheta\frac{\partial Y_l^m}{\partial\vartheta}\right) + \frac{1}{\sin^2\vartheta}\frac{\partial^2 Y_l^m}{\partial\varphi^2}\right\}. \tag{9.136c}$$

Equation (9.136c) can be satisfied if the expression on the left-hand side depending only on r and expression on the right-hand side depending only on ϑ and φ are equal to a constant, i.e., both sides being independent of each other are equal to the same constant. From the partial differential equation two ordinary differential equations follow. If the constant is chosen equal to $l(l+1)$, then the so-called *radial equation* results depending only on r and the potential $V(r)$:

$$\frac{1}{R_lr^2}\frac{d}{dr}\left(r^2\frac{dR_l}{dr}\right) + \frac{2m}{\hbar^2}\left[E - V(r) - \frac{l(l+1)\hbar^2}{2mr^2}\right] = 0. \tag{9.136d}$$

To find a solution for the angle-dependent part also in the separated form

$$Y_l^m(\vartheta,\varphi) = \Theta(\vartheta)\Phi(\varphi) \tag{9.136e}$$

one substitutes (9.136e) into (9.136c) giving

$$\sin^2\vartheta\left\{\frac{1}{\Theta\sin\vartheta}\frac{d}{d\vartheta}\left(\sin\vartheta\frac{d\Theta}{d\vartheta}\right) + l(l+1)\right\} = -\frac{1}{\Phi}\frac{d^2\Phi}{d\varphi^2}. \tag{9.136f}$$

If the separation constant is chosen as m^2 in a reasonable way, then the so-called *polar equation* is

$$\frac{1}{\Theta\sin\vartheta}\frac{d}{d\vartheta}\left(\sin\vartheta\frac{d\Theta}{d\vartheta}\right) + l(l+1) - \frac{m^2}{\sin^2\vartheta} = 0 \tag{9.136g}$$

and the *azimuthal equation* is

$$\frac{d^2\Phi}{d\varphi^2} + m^2\Phi = 0. \tag{9.136h}$$

Both equations are potential-independent, so they are valid for every central symmetric potential. There are three requirements for (9.136a): It should tend to zero for $r \to \infty$, it should be one-valued and quadratically integrable on the surface of the sphere.

3. Solution of the Radial Equation

Beside the potential $V(r)$ the radial equation (9.136d) also contains the separation constant $l(l+1)$. Substituting

$$u_l(r) = r \cdot R_l(r), \tag{9.137a}$$

since the square of the function $u_l(r)$ gives the last required probability $|u_l(r)|^2dr = |R_l(r)|^2r^2dr$ of the presence of the particle in a spherical shell between r and $r + dr$. The substitution leads to the one-dimensional Schroedinger equation

$$\frac{d^2u_l(r)}{dr^2} + \frac{2m}{\hbar^2}\left[E - V(r) - \frac{l(l+1)\hbar^2}{2mr^2}\right]u_l(r) = 0. \tag{9.137b}$$

This one contains the effective potential

$$V_{\text{eff}} = V(r) + V_l(l), \tag{9.137c}$$

which has two parts. The rotation energy

$$V_l(l) = V_{\text{rot}}(l) = \frac{l(l+1)\hbar^2}{2mr^2} \tag{9.137d}$$

is called the *centrifugal potential*.
The physical meaning of l as the *orbital angular momentum* follows from analogy with the classical rotation energy

$$E_{rot} = \frac{1}{2}\Theta\vec{\omega}^2 = \frac{(\Theta\vec{\omega})^2}{2\Theta} = \frac{\vec{l}^{\,2}}{2\Theta} = \frac{\vec{l}^{\,2}}{2mr^2} \tag{9.137e}$$

a rotating particle with moment of inertia $\Theta = \mu r^2$ and orbital angular momentum $\vec{l} = \Theta\vec{\omega}$:

$$\vec{l}^{\,2} = l(l+1)\hbar^2, \quad \left|\vec{l}^{\,2}\right| = \hbar\sqrt{l(l+1)}\,. \tag{9.137f}$$

4. Solution of the polar equation

The polar equation (9.136g), containing both separation constants $l(l+1)$ and m^2, is a Legendre differential equation (9.60a), p. 565. Its solution is denoted by $\Theta_l^m(\vartheta)$, and it can be determined by a power series expansion. Finite, single-valued and continuous solutions exist only for $l(l+1) = 0, 2, 6, 12, \ldots$. One gets for l and m:

$$l = 0, 1, 2, \ldots, \quad |m| \le l. \tag{9.138a}$$

So, m can take the $(2l+1)$ values

$$-l,\ (-l+1),\ (-l+2), \ldots,\ (l-2),\ (l-1),\ l. \tag{9.138b}$$

For $m \neq 0$ one gets the *corresponding Legendre polynomials*, which are defined in the following way:

$$P_l^m(\cos\vartheta) = \frac{(-1)^m}{2^l l!}(1-\cos^2\vartheta)^{m/2}\frac{d^{l+m}(\cos^2\vartheta - 1)^l}{(d\cos\vartheta)^{l+m}}. \tag{9.138c}$$

As a special case ($l = 0,\ m = n,\ \cos\vartheta = x$) follow the Legendre polynomials of the first kind (9.60c), p. 566. Their normalization results in the equation

$$\Theta_l^m(\vartheta) = \sqrt{\frac{2l+1}{2}\cdot\frac{(l-m)!}{(l+m)!}}\cdot P_l^m(\cos\vartheta) = N_l^m P_l^m(\cos\vartheta). \tag{9.138d}$$

5. Solution of the Azimuthal Equation

Since the motion of the particle in the potential field $V(r)$ is independent of the azimuthal angle even in the case of the physical assignment of a space direction, e.g., by a magnetic field, the general solution $\Phi = \alpha e^{\mathrm{i}m\varphi} + \beta e^{-\mathrm{i}m\varphi}$ can be specified by fixing

$$\Phi_m(\varphi) = Ae^{\pm\mathrm{i}m\varphi}, \tag{9.139a}$$

because in this case $|\Phi_m|^2$ is independent of φ. The requirement for uniqueness is

$$\Phi_m(\varphi + 2\pi) = \Phi_m(\varphi), \tag{9.139b}$$

so m can take on only the values $0, \pm1, \pm2, \ldots$.
It follows from the normalization

$$\int_0^{2\pi} |\Phi|^2\, d\varphi = 1 = |A|^2 \int_0^{2\pi} d\varphi = 2\pi|A|^2 \tag{9.139c}$$

that

$$\Phi_m(\varphi) = \frac{1}{\sqrt{2\pi}}e^{\mathrm{i}m\varphi} \quad (m = 0, \pm1, \pm2, \ldots). \tag{9.139d}$$

The quantum number m is called the *magnetic quantum number*.

6. Complete Solution for the Dependency of the Angles

In accordance with (9.136e), the solutions for the polar and the azimuthal equations should be multiplied by each other:

$$Y_l^m(\vartheta,\varphi) = \Theta(\vartheta)\,\Phi(\varphi) = \frac{1}{\sqrt{2\pi}} N_l^m P_l^m(\cos\vartheta) e^{im\varphi}. \tag{9.140a}$$

The functions $Y_l^m(\vartheta,\varphi)$ are the so-called *surface spherical harmonics*.

When the radius vector $\vec{\mathbf{r}}$ is reflected with respect to the origin ($\vec{\mathbf{r}} \to -\vec{\mathbf{r}}$), the angle ϑ becomes $\pi - \vartheta$ and φ becomes $\varphi + \pi$, so the sign of Y_l^m may change:

$$Y_l^m(\pi-\vartheta,\varphi+\pi) = (-1)^l Y_l^m(\vartheta,\varphi). \tag{9.140b}$$

Then for the *parity* of the considered wave function holds

$$P = (-1)^l. \tag{9.141a}$$

7. Parity

The *parity* property serves the characterization of the behavior of the wave function under *space inversion* $\vec{\mathbf{r}} \to -\vec{\mathbf{r}}$ (see 4.3.5.1, **1.**, p. 287). It is performed by the inversion or parity operator $\mathbf{P}$: $\mathbf{P}\Psi(\vec{\mathbf{r}},t) = \Psi(-\vec{\mathbf{r}},t)$. Denoting the eigenvalue of the operator by P, then applying $\mathbf{P}$ twice it must yield $\mathbf{P}\,\mathbf{P}\,\Psi(\vec{\mathbf{r}},t) = P\,P\,\Psi(\vec{\mathbf{r}},t) = \Psi(\vec{\mathbf{r}},t)$, the original wave function. So:

$$P^2 = 1, \quad P = \pm 1. \tag{9.141b}$$

It is called an *even wave function* if its sign does not change under space inversion, and it is called an *odd wave function* if its sign changes.

9.2.4.7 Linear Harmonic Oscillator

1. Formulation of the Problem

Harmonic oscillation occurs when the drag forces in the oscillator satisfy Hooke's law $F = -kx$. For the frequency of the oscillation, for the frequency of the oscillation circuit and for the potential energy the following formulas are valid:

$$\nu = \frac{1}{2\pi}\sqrt{\frac{k}{m}}, \quad (9.142a) \qquad \omega = \sqrt{\frac{k}{m}}, \quad (9.142b) \qquad E_{\text{pot}} = \frac{1}{2}kx^2 = \frac{\omega^2}{2}x^2. \quad (9.142c)$$

Substituting into (9.114a), the Schroedinger equation becomes

$$\frac{d^2\Psi}{dx^2} + \frac{2m}{\hbar^2}\left[E - \frac{\omega^2}{2}mx^2\right]\Psi = 0. \tag{9.143a}$$

With the substitutions

$$y = x\sqrt{\frac{m\omega}{\hbar}}, \quad (9.143b) \qquad \lambda = \frac{2E}{\hbar\omega}, \quad (9.143c)$$

where λ is a parameter and not the wavelength, (9.143a) can be transformed into the simpler form of the Weber differential equation

$$\frac{d^2\Psi}{dy^2} + (\lambda - y^2)\Psi = 0. \tag{9.143d}$$

2. Solution

A solution of the Weber differential equation can be got in the form

$$\Psi(y) = e^{-y^2/2}H(y). \tag{9.144a}$$

Differentiation shows that

$$\frac{d^2\Psi}{dy^2} = e^{-y^2/2}\left[\frac{d^2H}{dy^2} - 2y\frac{dH}{dy} + (y^2-1)H\right]. \tag{9.144b}$$

Substitution into (9.143d) yields

$$\frac{d^2H}{dy^2} - 2y\frac{dH}{dy} + (\lambda - 1)H = 0. \tag{9.144c}$$

The determination of a solution is convenient in the form of a series:

$$H = \sum_{i=0}^{\infty} a_i y^i \quad \text{with} \quad \frac{dH}{dy} = \sum_{i=1}^{\infty} i a_i y^{i-1}, \quad \frac{d^2H}{dy^2} = \sum_{i=2}^{\infty} i(i-1)a_i y^{i-2}. \tag{9.145a}$$

Substitution of (9.145a) into (9.144c) results in

$$\sum_{i=2}^{\infty} i(i-1)a_i y^{i-2} - \sum_{i=1}^{\infty} 2i a_i y^i + \sum_{i=0}^{\infty} i(\lambda-1)a_i y^i = 0. \tag{9.145b}$$

Comparing the coefficients of y^j leads to the recursion formula

$$(j+2)(j+1)a_{j+2} = [2j - (\lambda - 1)]a_j \quad (j = 0, 1, 2, \ldots). \tag{9.145c}$$

The coefficients a_j for even powers of y begin from a_0, the coefficients for odd powers begin from a_1. So, a_0 and a_1 can be chosen arbitrarily.

3. Physical Solutions

The determination of the probability of the presence of a particle in the different states can be performed by a quadratically integrable wave function $\Psi(x)$ and by an eigenfunction which has physical meaning, i.e., normalizable and for large values of y it tends to zero.

The exponential function $\exp(-y^2/2)$ in (9.144a) guarantees that the solution $\Psi(y)$ tends to zero for $y \to \infty$ if the function $H(y)$ is a polynomial. To get a polynomial, the coefficients a_j in (9.145a), starting from a certain n, must vanish for every $j > n$: $a_n \neq 0$, $a_{n+1} = a_{n+2} = a_{n+3} = \ldots = 0$. The recursion formula (9.145c) with $j = n$ is

$$a_{n+2} = \frac{2n - (\lambda - 1)}{(n+2)(n+1)} a_n. \tag{9.146a}$$

$a_{n+2} = 0$ can be satisfied for $a_n \neq 0$ only if

$$2n - (\lambda - 1) = 0, \quad \lambda = \frac{2E}{\hbar\omega} = 2n + 1. \tag{9.146b}$$

The coefficients $a_{n+2}, a_{n+4}, \ldots$ vanish for this choice of λ. Also $a_{n-1} = 0$ must hold to make the coefficients $a_{n+1}, a_{n+3}, \ldots$ equal to zero.

One gets the *Hermite polynomials* from the second defining equation (see 9.1.2.6, **6.**, p. 568) for the special choice of $a_n = 2^n$, $a_{n-1} = 0$. The first six of them are:

$$\begin{aligned} &H_0(y) = 1, && H_3(y) = -12y + 8y^3, \\ &H_1(y) = 2y, && H_4(y) = 12 - 48y^2 + 16y^4, \\ &H_2(y) = -2 + 4y^2, && H_5(y) = 120y - 160y^3 + 32y^5. \end{aligned} \tag{9.146c}$$

The solution $\Psi(y)$ for the *vibration quantum number* n is

$$\Psi_n = N_n e^{-y^2/2} H_n(y), \tag{9.147a}$$

where N_n is the normalizing factor. One gets it from the normalization condition $\int \Psi_n{}^2\, dy = 1$ as

$$N_n{}^2 = \frac{1}{2^n n!}\sqrt{\frac{\alpha}{\pi}} \quad \text{with} \quad \sqrt{\alpha} = \frac{y}{x} = \sqrt{\frac{m\omega}{\hbar}} \quad \text{(see (9.143b), p. 601).} \tag{9.147b}$$

From the terminating condition of the series (9.143c)

$$E_n = \hbar\omega\left(n + \frac{1}{2}\right) \quad (n = 0, 1, 2, \ldots) \tag{9.147c}$$

follows for the eigenvalues of the vibration energy. The spectrum of the energy levels is equidistant. The summand $+1/2$ in the parentheses means that in contrast to the classical case the quantum mechanical oscillator has energy even in the deepest energetic level with $n = 0$, which is known as the *zero-point vibration energy.*

Fig. 9.21 shows a graphical representation of the equidistant spectra of the energy states, the corresponding wave functions from Ψ_0 to Ψ_5 and also the function of the potential energy (9.142c). The points of the parabola of the potential energy represent the reversal points of the classical oscillator, which are calculated from the energy $E = \frac{1}{2}m\omega^2 a^2$ as the amplitude $a = \frac{1}{\omega}\sqrt{\frac{2E}{m}}$. The quantum mechanical probability of finding a particle in the interval $(x, x+dx)$ is given by $dw_{qu} = |\Psi(x)|^2\,dx$. It is different from zero also outside of these points. So for, e.g., $n = 1$, hence for $E = (3/2)\hbar\omega$, according to $dw_{qu} = 2\sqrt{\frac{\lambda}{\pi}}e^{-\lambda x^2}\,dx$, the maximum of the probability of presence is at

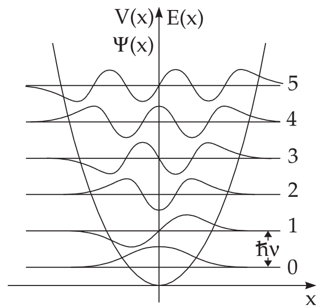


Figure 9.21

$$x_{max,qu} = \frac{\pm 1}{\sqrt{\lambda}} = \pm\sqrt{\frac{\hbar}{m\omega}}. \tag{9.147d}$$

For a corresponding classical oscillator, this is

$$x_{max,kl} = \pm a = \pm\sqrt{\frac{2E}{m\omega^2}} = \pm\sqrt{\frac{3\hbar}{m\omega}}. \tag{9.147e}$$

For large quantum numbers n the quantum mechanical probability density function approaches the classical one in its mean value.

9.2.5 Non-Linear Partial Differential Equations: Solitons, Periodic Patterns and Chaos

9.2.5.1 Formulation of the Physical-Mathematical Problem

1. Notion of Solitons

Solitons, also called solitary waves, from the viewpoint of physics, are pulses, or also localized disturbances of a non-linear medium or field; the energy related to such propagating pulses or disturbances is concentrated in a narrow spatial region. They occur:

- in solids, e.g., in inharmonic lattices, in Josephson contacts, in glass fibres and in quasi-one-dimensional conductors,
- in fluids as surface waves or spin waves,
- in plasmas as Langmuir solitons,
- in linear molecules,
- in classical and quantum field theory.

Solitons have both particle and wave properties; they are localized during their evolution, and the domain of the localization, or the point around which the wave is localized, travels as a free particle; in

particular it can also be at rest. A soliton has a permanent wave structure: based on a balance between nonlinearity and dispersion, the form of this structure does not change.
Mathematically, solitons are special solutions of certain non-linear partial differential equations occurring in physics, engineering and applied mathematics. Their special features are the absence of any dissipation and also that the non-linear terms cannot be handled by perturbation theory. Dissipative solitons are in non-conservative systems.

2. Important Examples of Equations with Soliton Solutions

a) Korteweg de Vries (KdV) Equation $$u_t + 6uu_x + u_{xxx} = 0, \tag{9.148}$$

b) Non-Linear Schroedinger (NLS) Equation $$\mathrm{i}\, u_t + u_{xx} \pm 2|u|^2 u = 0, \tag{9.149}$$

c) Sine-Gordon (SG) Equation $$u_{tt} - u_{xx} + \sin u = 0. \tag{9.150}$$

The subscripts x and t denote partial derivatives, e.g., $u_{xx} = \partial^2 u/\partial x^2$.
In these equations the one-dimensional case is considered, i.e., u has the form $u = u(x,t)$, where x is the spatial coordinate and t is the time. The equations are given in a scaled form, i.e., the two independent variables x and t are here dimensionless quantities. In practical applications, they must be multiplied by quantities having the corresponding dimensions and being characteristic of the given problem. The same holds for the velocity.

3. Interaction between Solitons

If two solitons, moving with different velocities, collide, they appear again after the interaction as if they had not collided. Every soliton asymptotically keeps its form and velocity; there is only a phase shift. Two solitons can interact without disturbing each other asymptotically. This is called an elastic interaction which is equivalent to the existence of an N-soliton solution, where N $(N = 1, 2, 3, \ldots)$ is the number of solitons. Solving an initial value problem with a given initial pulse $u(x,0)$ that disaggregate into solitons, the number of solitons does not depend on the shape of the pulse but on its total amount $\int_{-\infty}^{+\infty} u(x,0)\,dx$.

4. Periodic Patterns and Non-Linear Waves

Such non-linear phenomena occur in several classic dissipative systems (i.e. friction or damping systems), when an external impact or force is sufficiently large. E.g., if there is a layer of fluid in the gravitational field, and it is heated from below, the difference of temperature between the upper and lower surface corresponds to an external force. The higher temperature of the lower layer reduces its density and makes it lighter than the upper part, so the layering becomes unstable. At a sufficiently large temperature difference this unstable layering turns spontaneously into periodically arranged convection cells. It is called the *bifurcation* from the state of thermal conductivity (without convection) into the well ordered Rayleigh-Bénard convection. Taking away the external force results because of dissipation into damping of the waves (here the cellular convection). Strengthening of the external force drives the ordered convection into a turbulent convection and into chaos (see 17.3, p. 892). Also in chemical reactions similar phenomena can occur. Important examples for equations describing such phenomena are:

a) Ginsburg-Landau (GL) Equation $$u_t - u - (1 + \mathrm{i}b)u_{xx} + (1 + \mathrm{i}c)|u|^2 u = 0, \tag{9.151}$$

b) Kuramoto-Sivashinsky (KS) Equation $$u_t + u_{xx} + u_{xxxx} + u_x^2 = 0. \tag{9.152}$$

In contrast to the dissipation-less KdV, NLS, SG, equations, the equations (9.151) and (9.152) are non-linear dissipative equations, which have, besides spatiotemporal periodic solutions, also spatiotemporal disordered (chaotic) solutions. Appearance of spatiotemporal patterns or structures is characteristic which turn into chaos.

5. Dissipative Solitons

Solitary (isolated) wave phenomena in non-conservative systems often are called *dissipative solitons.* Contrary to the conservative systems, in which the solitons usually form families of solutions with at least one continuously changing parameter, dissipative solitons can be found at single points of the parameter space, at which a balance is formed between dispersion and nonlinearity from one side and

energy or particle flow and dissipation from the other side. This property leads to a special kind of stability of dissipative solitons, although they are not solutions of integrable wave equations. Dissipative solitons are described among others by the complex Ginsburg-Landau equation. They occur ,e.g., in non-linear optic cavitations, in optic semiconductor amplifiers and in reaction-diffusion systems (see also [9.16]).

6. Non-Linear Evolution Equations

Evolution equations describe the evolution of a physical quantity in time. Examples are the wave equation (see 9.2.3.2, p. 590), the heat equation (see 9.2.3.3, p. 591) and the Schroedinger equation (see 9.2.4.1, **1.**, p. 592). The solutions of the evolution equations are called *evolution functions*.

In contrast to linear evolution equations, the non-linear evolution equations (9.148), (9.149), and (9.150) contain non-linear terms like $u\partial u/\partial x$, $|u|^2 u$, $\sin u$ and u_x^2. These equations are (with the exception of (9.151)) parameter-free. From the viewpoint of physics non-linear evolution equations describe structure formations like solitons (dispersive structures) as well as periodic patterns and non-linear waves (dissipative structures).

9.2.5.2 Korteweg de Vries Equation (KdV)

1. Occurrence

The KdV equation is used in the discussion of
- surface waves in shallow water,
- inharmonic vibrations in non-linear lattices,
- problems of plasma physics and
- non-linear electric networks.

2. Equation and Solutions

The KdV equation for the evolution function u is

$$u_t + 6uu_x + u_{xxx} = 0. \tag{9.153}$$

It has the soliton solution

$$u(x,t) = \frac{v}{2\cosh^2\left[\frac{1}{2}\sqrt{v}(x - vt - \varphi)\right]}. \tag{9.154}$$

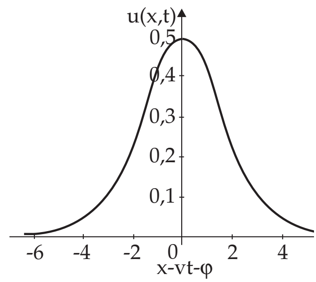


Figure 9.22

This KdV soliton is uniquely defined by the two dimensionless parameters v ($v > 0$) and φ. In **Fig. 9.22** $v = 1$ is chosen. A typical non-linear effect is that the velocity of the soliton v determines the amplitude and the width of the soliton: KdV solitons with larger amplitude and smaller width move faster than those with smaller amplitude and larger width (taller waves travel faster than shorter ones). The soliton phase φ describes the position of the maximum of the soliton at time $t = 0$.

Equation (9.153) also has N-solitons solutions. Such an N-soliton solution can be represented asymptotically for $t \to \pm\infty$ by the linear superposition of one-soliton solutions:

$$u(x,t) \sim \sum_{n=1}^{N} u_n(x,t). \tag{9.155}$$

Here every evolution function $u_n(x,t)$ is characterized by a velocity v_n and a phase $\varphi_n^{\pm}$. The initial phases φ_n^- before the interaction or collision differ from the final phases after the collision φ_n^+, while the velocities $v_1, v_2, \ldots, v_N$ have no changes, i.e., it is an elastic interaction.

For $N = 2$, (9.153) has a two-soliton solution. It cannot be represented for a finite time by a linear superposition, and with $k_n = \dfrac{1}{2}\sqrt{v_n}$ and $\alpha_n = \dfrac{1}{2}\sqrt{v_n}(x - v_n t - \varphi_n)$ $(n = 1, 2)$ it has the form:

$$u(x,t) = 8\frac{k_1^2 e^{\alpha_1} + k_2^2 e^{\alpha_2} + (k_1 - k_2)^2 e^{(\alpha_1+\alpha_2)}\left[2 + \frac{1}{(k_1+k_2)^2}\left(k_1^2 e^{\alpha_1} + k_2^2 e^{\alpha_2}\right)\right]}{\left[1 + e^{\alpha_1} + e^{\alpha_2} + \left(\frac{k_1 - k_2}{k_1 + k_2}\right)^2 e^{(\alpha_1+\alpha_2)}\right]^2}. \quad (9.156)$$

Equation (9.156) describes two non-interacting solitons for $t \to -\infty$ asymptotically with velocities $v_1 = 4k_1^2$ and $v_2 = 4k_2^2$, which transform after their mutual interaction again into two non-interacting solitons with the same velocities for $t \to +\infty$ asymptotically.
The non-linear evolution equation

$$w_t + 6(w_x)^2 + w_{xxx} = 0 \quad (9.157a)$$

where $w = \frac{F_x}{F}$ has the following properties:

a) For $F(x,t) = 1 + e^{\alpha}, \quad \alpha = \frac{1}{2}\sqrt{v}(x - vt - \varphi)$ (9.157b)

it has a soliton solution and

b) for $F(x,t) = 1 + e^{\alpha_1} + e^{\alpha_2} + \left(\frac{k_1 - k_2}{k_1 + k_2}\right)^2 e^{(\alpha_1+\alpha_2)}$ (9.157c)

it has a two-soliton solution. With $2w_x = u$ the KdV equation (9.153) follows from (9.157a). Equation (9.156) and the expression w following from (9.157c) are examples of a non-linear superposition.

If the term $+6uu_x$ is replaced by $-6uu_x$ in (9.153), then the right-hand side of (9.154) has to be multiplied by (-1). In this case the notation *antisoliton* is used.

9.2.5.3 Non-Linear Schroedinger Equation (NLS)

1. Occurrence

The NLS equation occurs

• in non-linear optics, where the refractive index n depends on the electric field strength $\vec{\mathbf{E}}$, as, e.g., for the Kerr effect, where $n(\vec{\mathbf{E}}) = n_0 + n_2|\vec{\mathbf{E}}|^2$ with $n_0, n_2 =$ constant holds, and
• in the hydrodynamics of self-gravitating discs which allow us to describe galactic spiral arms.

2. Equation and Solution

The NLS equation for the evolution function u and its solution are:

$$\mathrm{i}\, u_t + u_{xx} \pm 2|u|^2 u = 0, \quad (9.158) \qquad u(x,t) = 2\eta\frac{\exp\left(-\mathrm{i}[2\xi x + 4(\xi^2 - \eta^2)t - \chi]\right)}{\cosh[2\eta(x + 4\xi t - \varphi)]}. \quad (9.159)$$

Here $u(x,t)$ is complex. The NLS soliton is characterized by the four dimensionless parameters η, ξ, φ, and χ. The envelope of the wave packet moves with the velocity $v = -4\xi$; the phase velocity of the wave packet is $2(\eta^2 - \xi^2)/\xi$.

In contrast to the KdV soliton (9.154), the amplitude and the velocity can be chosen independently of each other. **Fig. 9.23** displays the real part of (9.159) with $\eta = 1/2$ and $\xi = 4$.

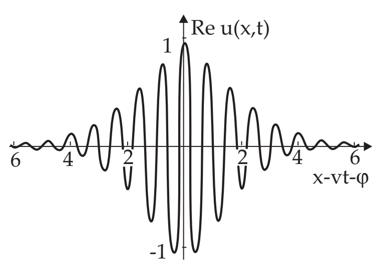


Figure 9.23

The solutions of the form (9.159) often are called *light solitons*; they solve the focusing NLS equation (9.158) for the case "+". The defocusing NLS equation (case "−") gives solitons, for which $|u|^2$ at the position of the

solitons is reduced in comparison to a constant background $|u(x \to \pm\infty)| = \eta$. Such *dark solitons* have the form

$$u(x,t) = \left(\mathrm{i}\frac{v}{2} + \sqrt{\eta^2 - \frac{v^2}{4}}\tanh\left[\sqrt{\eta^2 - \frac{v^2}{4}}(x - vt)\right]\right) \cdot \exp\left[-\mathrm{i}\left(2\eta^2 t + \chi\right)\right]. \qquad (9.160)$$

They depend on the three parameters η, v and χ and propagate with the velocity $v < 2\eta$ on a background with trivial (flat) phase (see [9.26], [9.23]).

A general solution has in addition a phase gradient, which can be interpreted as velocity c of the background, relative to which the soliton is moving. Then the solution has the following form:

$$u(x,t) = \left(\mathrm{i}\frac{v}{2} + \sqrt{\eta^2 - \frac{v^2}{4}}\right)\tanh\left[\sqrt{\eta^2 - \frac{v^2}{4}}(x - vt - ct)\right]\exp\left[-\mathrm{i}\left(2\eta^2 t + \chi - \frac{c}{2}x + \frac{c^2}{4}t\right)\right]. \qquad (9.161)$$

Beside of these exponential positioned soliton waves also periodic solutions of the NLS equation exist, which can be interpreted as wave packets of solitons. Such solutions can be find demanding stationarity and by integration of the remaining ordinary differential equation. Generally such solutions are elliptic Jacobian-functions (see 14.6.2, p. 763). Some relevant solutions see [9.17].

In the case of N interacting solitons, one can characterize them by $4N$ arbitrary chosen parameters: $\eta_n, \xi_n, \varphi_n, \chi_n \quad (n = 1, 2, \ldots, N)$.
If the solitons have different velocities, the N-soliton solution splits asymptotically for $t \to \pm\infty$ into a sum of N individual solitons of the form (9.159).

9.2.5.4 Sine-Gordon Equation (SG)

1. Occurrence

The SG equation is obtained from the Bloch equation for spatially inhomogeneous quantum mechanical two-level systems. It describes the propagation of
- ultra-short pulses in resonant laser media (self-induced transparency),
- the magnetic flux in large surface Josephson contacts, i.e., in tunnel contacts between two superconductors and
- spin waves in superfluid helium 3 (^{3}He).

The soliton solution of the SG equation can be illustrated by a mechanical model of pendula and springs. The evolution function goes continuously from 0 to a constant value c. The *SG solitons* are often called *kink solitons*. If the evolution function changes from the constant value c to 0, it describes a so-called *antikink soliton*. Walls of domain structures can be described with this type of solutions.

2. Equation and Solution

The SG equation for the evolution function u is

$$u_{tt} - u_{xx} + \sin u = 0. \qquad (9.162)$$

It has the following soliton solutions:

1. Kink Soliton

$$u(x,t) = 4\arctan e^{\gamma(x - x_0 - vt)}, \qquad (9.163)$$

where $\gamma = \dfrac{1}{\sqrt{1 - v^2}}$ and $-1 < v < +1$.

The kink soliton (9.163) for $v = 1/2$ is given in **Fig. 9.24**. The kink soliton is determined by two dimensionless parameters v and x_0. The velocity is independent of the amplitude. The time and the position derivatives are ordi-

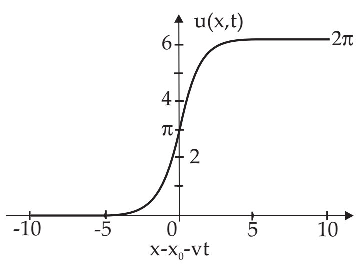


Figure 9.24

nary localized solitons:

$$-\frac{u_t}{v} = u_x = \frac{2\gamma}{\cosh\gamma(x-x_0-vt)}. \tag{9.164}$$

2. Antikink Soliton

$$u(x,t) = 4\arctan e^{-\gamma(x-x_0-vt)}. \tag{9.165}$$

3. Kink-Antikink Soliton One gets a static kink-antikink soliton from (9.163, 9.165) with $v = 0$:

$$u(x,t) = 4\arctan e^{\pm(x-x_0)}. \tag{9.166}$$

Further solutions of (9.162) are:

4. Kink-Kink Collision

$$u(x,t) = 4\arctan\left[v\frac{\sinh\gamma x}{\cosh\gamma vt}\right]. \tag{9.167}$$

5. Kink-Antikink Collision

$$u(x,t) = 4\arctan\left[\frac{1}{v}\frac{\sinh\gamma vt}{\cosh\gamma x}\right]. \tag{9.168}$$

6. Double or Breather Soliton, also called Kink-Antikink Doublet

$$u(x,t) = 4\arctan\left[\frac{\sqrt{1-\omega^2}}{\omega}\frac{\sin\omega t}{\cosh\sqrt{1-\omega^2}x}\right]. \tag{9.169}$$

Equation (9.169) represents a stationary wave, whose envelope is modulated by the frequency ω.

7. Local Periodic Kink Lattice

$$u(x,t) = 2\arcsin\left[\pm\mathrm{sn}\left(\frac{x-vt}{k\sqrt{1-v^2}}, k\right)\right] + \pi. \tag{9.170a}$$

The relation between the wavelength λ and the lattice constant k is

$$\lambda = 4K(k)k\sqrt{1-v^2}. \tag{9.170b}$$

For $k = 1$, i.e., for $\lambda \to \infty$, one gets

$$u(x,t) = 4\arctan e^{\pm\gamma(x-vt)}, \tag{9.170c}$$

which is the kink soliton (9.163) and the antikink soliton (9.165) again, with $x_0 = 0$.

Remark: $\mathrm{sn}\, x$ is a Jacobian elliptic function with parameter k and quarter-period K (see 14.6.2, p. 763):

$$\mathrm{sn}x = \sin\varphi(x,k), \tag{9.171a}$$

$$x = \int\limits_0^{\sin\varphi(x,k)} \frac{dq}{\sqrt{(1-q^2)(1-k^2q^2)}}, \quad (9.171b) \qquad K(k) = \int\limits_0^{\pi/2} \frac{d\Theta}{\sqrt{1-k^2\sin^2\Theta}}. \quad (9.171c)$$

Equation (9.171b) comes from (14.104a), p. 763, by the substitution of $\sin\psi = q$. The series expansion of the complete elliptic integral is given as equation (8.104), in 8.2.5, **7.**, p. 515.

9.2.5.5 Further Non-linear Evolution Equations with Soliton Solutions

1. Modified KdV Equation

$$u_t \pm 6u^2u_x + u_{xxx} = 0. \tag{9.172}$$

The even more general equation (9.173) has the soliton (9.174) as solution.

$$u_t + u^p u_x + u_{xxx} = 0\,, \quad (9.173) \qquad u(x,t) = \left[\frac{\frac{1}{2}|v|(p+1)(p+2)}{\cosh^2\left(\frac{1}{2}p\sqrt{|v|}(x - vt - \varphi)\right)}\right]^{\frac{1}{p}}. \quad (9.174)$$

2. Sinh-Gordon Equation

$$u_{tt} - u_{xx} + \sinh u = 0. \quad (9.175)$$

3. Boussinesq Equation

$$u_{xx} - u_{tt} + (u^2)_{xx} + u_{xxxx} = 0. \quad (9.176)$$

This equation occurs in the description of non-linear electric networks as a continuous approximation of the charge-voltage relation.

4. Hirota Equation

$$u_t + \mathrm{i}3\alpha|u|^2u_x + \beta u_{xx} + \mathrm{i}\sigma u_{xxx} + \delta|u|^2u = 0, \qquad \alpha\beta = \sigma\delta. \quad (9.177)$$

5. Burgers Equation

$$u_t - u_{xx} + uu_x = 0. \quad (9.178)$$

This equation occurs when modeling turbulence. With the Hopf-Cole transformation it is transformed into the diffusion equation, i.e., into a linear differential equation.

6. Kadomzev-Pedviashwili Equation

The equation

$$(u_t + 6uu_x + u_{xxx})_x = u_{yy} \quad (9.179a)$$

has the soliton

$$u(x,y,t) = 2\frac{\partial^2}{\partial x^2}\ln\left[\frac{1}{k^2} + \left|x + \mathrm{i}ky - 3k^2t\right|^2\right] \quad (9.179b)$$

as its solution. The equation (9.179a) is an example of a soliton equation with a higher number of independent variables, e.g., with two spatial variables.

Remark: The CD-ROM to the 7th, 8th and 9th German editions of this handbook (see [22.8]) contains more non-linear evolution equations. Furthermore there is shown the application of the Fourier transformation and of the inverse scattering theory to solve linear partial differential equations.

10 Calculus of Variations

10.1 Defining the Problem

1. Extremum of an Integral Expression

A very important problem of the differential calculus is to determine for which x values the given function $y(x)$ has extreme values. The calculus of variations discusses the following problem: For which functions has a certain integral, whose integrand depends also on the unknown function and its derivatives, an extremum value? The calculus of variations concerns itself with determining all the functions $y(x)$ for which the integral expression

$$I[y] = \int_a^b F(x, y(x), y'(x), \ldots, y^{(n)}(x))dx \tag{10.1}$$

has an extremum, if the functions $y(x)$ are from a previously given class of functions. Here, it ist possible to define some *boundary* and *side conditions* for $y(x)$ and for its derivatives.

2. Integral Expressions of Variational Calculus

There can also be several variables instead of x in (10.1). In this case, the occurring derivatives are partial derivatives and the integral in (10.1) is a multiple integral. In the calculus of variations, mainly the following types of integral expressions are discussed:

$$I[y] = \int_a^b F(x, y(x), y'(x))\, dx, \tag{10.2}$$

$$I[y_1, y_2, \ldots, y_n] = \int_a^b F(x, y_1(x), \ldots, y_n(x), y_1'(x), \ldots, y_n'(x))\, dx, \tag{10.3}$$

$$I[y] = \int_a^b F(x, y(x), y'(x), \ldots, y^{(n)}(x))\, dx, \tag{10.4}$$

$$I[u] = \iint_\Omega F(x, y, u, u_x, u_y)\, dx\, dy. \tag{10.5}$$

Here the unknown function is $u = u(x, y)$, and Ω represents a plane domain of integration.

$$I[u] = \iiint_R F(x, y, z, u, u_x, u_y, u_z)\, dx\, dy\, dz. \tag{10.6}$$

The unknown function is $u = u(x, y, z)$, and R represents a space region of integration. Additionally, boundary values can be given for the solution of a variational problem, at the endpoints of the interval a and b in the one-dimensional case, and at the boundary of the domain of integration Ω in the two-dimensional case. Besides, various further side conditions can be defined, e.g., in integral form or as a differential equation.

A variational problem is called *first-order* or *higher-order* depending whether the integrand F contains only the first derivative y' or higher derivatives $y^{(n)}$ $(n > 1)$ of the function y.

3. Parametric Representation of the Variational Problem

A variational problem can also be posed in *parametric form*. Considering a curve in parametric form $x = x(t),\ y = y(t)\ \ (\alpha \le t \le \beta)$, then, e.g., the integral expression (10.2) has the form

$$I[x, y] = \int_\alpha^\beta F(x(t), y(t), \dot{x}(t), \dot{y}(t))\, dt. \tag{10.7}$$

I.N. Bronshtein et al., *Handbook of Mathematics*,
DOI 10.1007/978-3-662-46221-8_10

10.2 Historical Problems

10.2.1 Isoperimetric Problem

The *general isoperimetric problem* is to determine the plane region with the largest area among the plane regions with a given perimeter. The solution of this problem, a circle with a given perimeter, originates from queen Dido, who was allowed, as legend has it, to take such an area for the foundation of Carthago which she could be surround by one bull's leather. She cut the leather into fine stripes, and formed a circle with them.

A special case of the isoperimetric problem is to find the equation of the curve $y = y(x)$ in a Cartesian coordinate system connecting the points $A(a, 0)$ and $B(b, 0)$ and having the given length l, for which the area determined by the line segment $\overline{AB}$ and the curve is the largest possible **(Fig. 10.1)**. For the mathematical formalization a once continuously differentiable function $y(x)$ is to be determined, such that

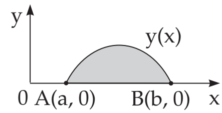


Figure 10.1

$$I[y] = \int_a^b y(x)\,dx = \max \tag{10.8a}$$

holds, where the side condition (10.8b) and the boundary conditions (10.8c) are satisfied:

$$G[y] = \int_a^b \sqrt{1 + y'^2(x)}\,dx = l \qquad (10.8b) \qquad\qquad y(a) = y(b) = 0 \qquad (10.8c)$$

10.2.2 Brachistochrone Problem

The brachistochrone problem was formulated in 1696 by J. Bernoulli, and it is the following: A point mass descends from the point $P_0(x_0, y_0)$ to the origin in the vertical plane x, y only under the influence of gravity. The curve $y = y(x)$ along which the point reaches the origin in the shortest possible time from P_0 **(Fig. 10.2)** is to be determined. With the formula for the time of fall T, (see 11.5.1, p. 648), the following mathematical description is possible to determine a once continuously differentiable function $y = y(x)$ for which

$$T[y] = \int_0^{x_0} \frac{\sqrt{1 + y'^2}}{\sqrt{2g(y_0 - y)}}\,dx = \min \tag{10.9}$$

holds, (g is the acceleration due to gravity). The boundary value conditions are

$$y(0) = 0, \qquad y(x_0) = y_0. \tag{10.10}$$

For $x = x_0$ in (10.9) there is a singularity.

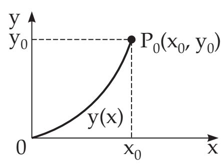


Figure 10.2

10.3 Variational Problems of One Variable

10.3.1 Simple Variational Problems and Extremal Curves

A *simple variational problem* is to determine the extreme value of the integral expression given in the form

$$I[y] = \int_a^b F(x, y(x), y'(x))\,dx, \tag{10.11}$$

where $y(x)$ is a twice continuously differentiable function satisfying the boundary conditions $y(a) = A$ and $y(b) = B$. The values a, b and A, B, and the function F are given.
The integral expression (10.11) is an example of a so-called *functional*. A functional assigns a real number to every function $y(x)$ from a certain class of functions.
If the functional $I[y]$ in (10.11) takes, e.g., its relative maximum for a function $y_0(x)$, then

$$I[y_0] \geq I[y] \tag{10.12}$$

for every twice continuously differentiable function y satisfying the boundary conditions. The curve $y = y_0(x)$ is called an *extremal curve*. Sometimes all the solutions of the Euler differential equation of the variational calculus are called extremal curves.

10.3.2 Euler Differential Equation of the Variational Calculus

A necessary condition for the solution of the variational problem can be constructed by the help of an *auxiliary curve* or *comparable curve* for the extremal $y_0(x)$ characterized by (10.12)

$$y(x) = y_0(x) + \epsilon\,\eta(x) \tag{10.13}$$

with a twice continuously differentiable function $\eta(x)$ satisfying the special boundary conditions $\eta(a) = \eta(b) = 0$; ϵ is a real parameter. Substituting (10.13) into (10.11) there is a function depending on ϵ instead of the functional $I[y]$

$$I(\epsilon) = \int\limits_a^b F(x, y_0 + \epsilon\,\eta, y_0' + \epsilon\,\eta')\,dx\,. \tag{10.14}$$

The functional $I[y]$ has an extreme value for $y_0(x)$ if the function $I(\epsilon)$, as a function of ϵ, has an extreme value for $\epsilon = 0$. Deducing the variational problem to an extreme value problem with the necessary condition

$$\frac{dI}{d\epsilon} = 0 \quad \text{for} \quad \epsilon = 0 \tag{10.15}$$

and supposing that the function F, as a function of three independent variables, is differentiable as many times as needed, by the help of the Taylor expansion (see 7.3.3.3, p. 471) follows

$$I(\epsilon) = \int\limits_a^b \left[F(x, y_0, y_0') + \frac{\partial F}{\partial y}(x, y_0, y_0')\epsilon\,\eta + \frac{\partial F}{\partial y'}(x, y_0, y_0')\epsilon\,\eta' + O(\epsilon^2) \right] dx. \tag{10.16}$$

The necessary condition (10.15) results in the equation

$$\int\limits_a^b \eta \frac{\partial F}{\partial y}\,dx + \int\limits_a^b \eta' \frac{\partial F}{\partial y'}\,dx = 0. \tag{10.17}$$

Partial integration of this equation and considering the boundary conditions for $\eta(x)$, gives

$$\int\limits_a^b \eta \left(\frac{\partial F}{\partial y} - \frac{d}{dx}\left(\frac{\partial F}{\partial y'}\right)\right) dx = 0. \tag{10.18}$$

From the assumption of continuity and because the integral in (10.18) must vanish for any considerable $\eta(x)$,

$$\frac{\partial F}{\partial y} - \frac{d}{dx}\left(\frac{\partial F}{\partial y'}\right) = 0 \tag{10.19}$$

must hold. The equation (10.19) gives a *necessary condition for the simple variational problem* and it is called the *Euler differential equation of the calculus of variations*. The differential equation (10.19)

can be written in the form

$$\frac{\partial F}{\partial y} - \frac{\partial^2 F}{\partial x \partial y'} - \frac{\partial^2 F}{\partial y \partial y'} y' - \frac{\partial^2 F}{\partial y'^2} y'' = 0. \tag{10.20}$$

It is an ordinary second-order differential equation if $F_{y'y'} \neq 0$ holds.
The Euler differential equation has a simpler form in the following special cases:
Case 1: $F(x, y, y') = F(y')$, i.e., x and y do not appear explicitly. Then instead of (10.19) holds

$$\frac{\partial F}{\partial y} = 0 \quad (10.21a) \qquad \text{and} \qquad \frac{d}{dx}\left(\frac{\partial F}{\partial y'}\right) = 0. \quad (10.21b)$$

Case 2: $F(x, y, y') = F(y, y')$, i.e., x does not appear explicitly. From

$$\frac{d}{dx}\left(F - y'\frac{\partial F}{\partial y'}\right) = \frac{\partial F}{\partial y} y' + \frac{\partial F}{\partial y'} y'' - y''\frac{\partial F}{\partial y'} - y'\frac{d}{dx}\left(\frac{\partial F}{\partial y'}\right) = y'\left(\frac{\partial F}{\partial y} - \frac{d}{dx}\left(\frac{\partial F}{\partial y'}\right)\right) \tag{10.22a}$$

and because of (10.19) follows

$$\frac{d}{dx}\left(F - y'\frac{\partial F}{\partial y'}\right) = 0, \quad (10.22b) \qquad \text{i.e.,} \qquad F - y'\frac{\partial F}{\partial y'} = c \quad (c \text{ const}) \quad (10.22c)$$

as a necessary condition for the solution of the simple variational problem in the case $F = F(y, y')$.
■ **A:** The functional to determine the shortest curve connecting the points $P_1(a, A)$ and $P_2(b, B)$ in the x, y plane is:

$$I[y] = \int_a^b \sqrt{1 + y'^2}\, dx = \min. \tag{10.23a}$$

It follows from (10.21b) for $F = F(y') = \sqrt{1 + y'^2}$ that

$$\frac{d}{dx}\left(\frac{\partial F}{\partial y'}\right) = \frac{y''}{\left(\sqrt{1 + y'^2}\right)^3} = 0, \tag{10.23b}$$

so $y'' = 0$, i.e., the shortest curve is the straight line.
■ **B:** Connecting the points $P_1(a, A)$ and $P_2(b, B)$ by a curve $y(x)$, and rotating it around the x-axis the surface area is

$$I[y] = 2\pi \int_a^b y\sqrt{1 + y'^2}\, dx. \tag{10.24a}$$

Which curve $y(x)$ creates the smallest surface area? From (10.22c) with $F = F(y, y') = 2\pi y\sqrt{1 + y'^2}$ follows $y = \frac{c}{2\pi}\sqrt{1 + y'^2}$ or $y'^2 = \left(\frac{y}{c_1}\right)^2 - 1$, where $c_1 = \frac{c}{2\pi}$. This differential equation is separable (see 9.2.2.3, **1.**, p. 579), and its solution is

$$y = c_1 \cosh\left(\frac{x}{c_1} + c_2\right) \qquad (c_1,\ c_2 \text{ const}), \tag{10.24b}$$

the equation of the so-called *catenary curve* (see 2.15.1, p. 107). The constants c_1 and c_2 can be determined from the boundary values $y(a) = A$ and $y(b) = B$. So, it is to solve a system of non-linear

equations (see 19.2.2, p. 961), which cannot be solved for every boundary value.

10.3.3 Variational Problems with Side Conditions

These problems are usually isoperimetric problems (see 10.2.1, p. 611): The simple variational problem (see 10.2.1, p. 611), given by the functional (10.11), is completed by a further side condition in the form

$$\int_a^b G(x, y(x), y'(x))\, dx = l \quad (l \text{ const}) \tag{10.25}$$

where the constant l and the function G are given. A method to solve this problem is given by Lagrange (extreme values with side conditions in equation form, see 6.2.5.6, p. 456). Considering the expression

$$H(x, y(x), y'(x), \lambda) = F(x, y(x), y'(x)) + \lambda G(x, y(x), y'(x)), \tag{10.26}$$

where λ is a parameter, and solving the problem

$$\int_a^b H(x, y(x), y'(x), \lambda) dx = \text{extreme!}, \tag{10.27}$$

as an extreme value problem without side condition. The corresponding Euler differential equation is:

$$\frac{\partial H}{\partial y} - \frac{d}{dx}\left(\frac{\partial H}{\partial y'}\right) = 0. \tag{10.28}$$

The solution $y = y(x, \lambda)$ depends on the parameter λ, which must be determined by substituting $y(x, \lambda)$ into the side condition (10.25).

■ For the isoperimetric problem 10.2.1, p. 611, one gets

$$H(x, y(x), y'(x), \lambda) = y + \lambda\sqrt{1 + y'^2}\,. \tag{10.29a}$$

Because the variable x does not appear in H, instead of the Euler differential equation (10.28), analogously to (10.22c), one gets the differential equation

$$y + \lambda\sqrt{1 + y'^2} - \frac{\lambda y'^2}{\sqrt{1 + y'^2}} = c_1 \quad \text{or} \quad y'^2 = \frac{\sqrt{\lambda^2 - (c_1 - y)^2}}{c_1 - y} \quad (c_1 \text{ const}), \tag{10.29b}$$

whose solution is a family of circles

$$(x - c_2)^2 + (y - c_1)^2 = \lambda^2 \quad (c_1, c_2, \lambda \text{ const}). \tag{10.29c}$$

The values c_1, c_2 and λ are determined from the conditions $y(a) = 0,\ y(b) = 0$ and from the requirement that the arclength between A and B should be l. The result is a non-linear equation for λ, which should be solved by an appropriate iterative method.

10.3.4 Variational Problems with Higher-Order Derivatives

There are two types of problems to be considered here.

1. $F = F(x, y, y', y'')$

The variational problem is:

$$I[y] = \int_a^b F(x, y, y', y'')\, dx = \text{extreme!} \tag{10.30a}$$

with the boundary values

$$y(a) = A, \quad y(b) = B, \quad y'(a) = A', \quad y'(b) = B', \tag{10.30b}$$

where the numbers a, b, A, B, A', and B', and the function F are given. Similarly as in 10.3.2, p. 612, introducing comparable curves $y(x) = y_0(x) + \epsilon\,\eta(x)$ with $\eta(a) = \eta(b) = \eta'(a) = \eta'(b) = 0$, yields the *Euler differential equation*

$$\frac{\partial F}{\partial y} - \frac{d}{dx}\left(\frac{\partial F}{\partial y'}\right) + \frac{d^2}{dx^2}\left(\frac{\partial F}{\partial y''}\right) = 0 \tag{10.31}$$

as a necessary condition for the solution of the variational problem (10.30a). The differential equation (10.31) represents a fourth-order differential equation. Its general solution contains four arbitrary constants which can be determined by the boundary values (10.30b).

■ Consider the problem

$$I[y] = \int_0^1 (y''^2 - \alpha y'^2 - \beta y^2)\,dx = \text{extreme!} \tag{10.32a}$$

with the given constants α and β for $F = F(y, y', y'') = y''^2 - \alpha y'^2 - \beta y^2$. Then: $F_y = -2\beta y$, $F_{y'} = -2\alpha y'$, $F_{y''} = 2y''$, $\frac{d}{dx}(F_{y'}) = -2\alpha y''$, $\frac{d^2}{dx^2}(F_{y''}) = 2y^{(4)}$, and the Euler differential equation is

$$y^{(4)} + \alpha y'' - \beta y = 0. \tag{10.32b}$$

This is a fourth-order linear differential equation with constant coefficients (see 9.1.2.3, p. 553).

2. $F = F(x, y, y', \dots, y^{(n)})$

In this general case, when the functional $I[y]$ of the variational problem depends on the derivatives of the unknown function y up to order n $(n \ge 1)$, the corresponding Euler differential equation is

$$\frac{\partial F}{\partial y} - \frac{d}{dx}\left(\frac{\partial F}{\partial y'}\right) + \frac{d^2}{dx^2}\left(\frac{\partial F}{\partial y''}\right) - \cdots + (-1)^n \frac{d^n}{dx^n}\left(\frac{\partial F}{\partial y^{(n)}}\right) = 0, \tag{10.33}$$

whose solution should satisfy the boundary conditions analogously to (10.30b) up to order $n-1$.

10.3.5 Variational Problem with Several Unknown Functions

Suppose the functional of the variational problem has the form

$$I[y_1, y_2, \dots, y_n] = \int_a^b F(x, y_1, y_2, \dots, y_n, y_1', y_2', \dots, y_n')\,dx, \tag{10.34}$$

where the unknown functions $y_1(x), y_2(x), \dots, y_n(x)$ should take given values at $x = a$ and $x = b$. Considering n twice continuously differentiable comparable functions

$$y_i(x) = y_{i0}(x) + \epsilon_i \eta_i(x) \quad (i = 1, 2, \dots, n), \tag{10.35}$$

where the functions $\eta_i(x)$ should vanish at the endpoints, (10.34) becomes $I(\epsilon_1, \epsilon_2, \dots, \epsilon_n)$ with (10.35). From the necessary conditions

$$\frac{\partial I}{\partial \epsilon_i} = 0 \quad (i = 1, 2, \dots, n) \tag{10.36}$$

for the extreme values of a function of several variables, follow the n Euler differential equations

$$\frac{\partial F}{\partial y_1} - \frac{d}{dx}\left(\frac{\partial F}{\partial y_1'}\right) = 0, \quad \frac{\partial F}{\partial y_2} - \frac{d}{dx}\left(\frac{\partial F}{\partial y_2'}\right) = 0, \quad \dots, \quad \frac{\partial F}{\partial y_n} - \frac{d}{dx}\left(\frac{\partial F}{\partial y_n'}\right) = 0, \tag{10.37}$$

whose solutions $y_1(x), y_2(x), \dots, y_n(x)$ must satisfy the given boundary conditions.

10.3.6 Variational Problems using Parametric Representation

For some variational problems it is useful to determine the extremal, not in the explicit form $y = y(x)$, but in the parametric form

$$x = x(t), \quad y = y(t) \quad (t_1 \le t \le t_2), \tag{10.38}$$

where t_1 and t_2 are the parameter values corresponding to the points (a, A) and (b, B). Then the simple variational problem (see 10.3.1, p. 611) is

$$I[x, y] = \int_{t_1}^{t_2} F(x(t), y(t), \dot{x}(t), \dot{y}(t))dt = \text{extreme!} \tag{10.39a}$$

with the boundary conditions

$$x(t_1) = a, \quad x(t_2) = b, \quad y(t_1) = A, \quad y(t_2) = B. \tag{10.39b}$$

Here $\dot{x}$ and $\dot{y}$ denote the derivatives of x and y with respect to the parameter t, as usual in the parametric representation.

The variational problem (10.39a) makes sense only if the value of the integral is independent of the parametric representation of the extremal curve. To ensure the integral in (10.39a) is independent of the parametric representation of the curve connecting the points (a, A) and (b, B), F must be a *positive homogeneous function* of first order, i.e.,

$$F(x, y, \mu\dot{x}, \mu\dot{y}) = \mu F(x, y, \dot{x}, \dot{y}) \quad (\mu > 0) \tag{10.40}$$

must hold.

Because the variational problem (10.39a) can be considered as a special case of (10.34), the corresponding Euler differential equations are

$$\frac{\partial F}{\partial x} - \frac{d}{dt}\left(\frac{\partial F}{\partial \dot{x}}\right) = 0, \quad \frac{\partial F}{\partial y} - \frac{d}{dt}\left(\frac{\partial F}{\partial \dot{y}}\right) = 0. \tag{10.41}$$

They are not independent of each other, but they are equivalent to the so-called *Weierstrass form* of the Euler differential equation:

$$\frac{\partial^2 F}{\partial x \partial \dot{y}} - \frac{\partial^2 F}{\partial \dot{x} \partial y} + M(\dot{x}\ddot{y} - \ddot{x}\dot{y}) = 0 \tag{10.42a}$$

with

$$M = \frac{1}{\dot{y}^2}\frac{\partial^2 F}{\partial \dot{x}^2} = -\frac{1}{\dot{x}\dot{y}}\frac{\partial^2 F}{\partial \dot{x}\partial \dot{y}} = \frac{1}{\dot{x}^2}\frac{\partial^2 F}{\partial \dot{y}^2}. \tag{10.42b}$$

Analogous to the calculation of the radius of curvature R of a curve given in parametric representation (see 3.6.1.1, **1.**, p. 243), the calculation of the *radius of curvature of the extremal curve* is to be made, considering (10.42a), with

$$R = \left|\frac{(\dot{x}^2 + \dot{y}^2)^{3/2}}{\dot{x}\ddot{y} - \ddot{x}\dot{y}}\right| = \left|\frac{M\,(\dot{x}^2 + \dot{y}^2)^{3/2}}{F_{\dot{x}y} - F_{x\dot{y}}}\right|. \tag{10.42c}$$

■ The isoperimetric problem (10.8a to 10.8c) (see 10.2.1, p. 611) has the form in parametric representation:

$$I[x, y] = \int_{t_1}^{t_2} y(t)\dot{x}(t)dt = \text{max!} \quad (10.43\text{a}) \qquad \text{with} \quad \int_{t_1}^{t_2} \sqrt{\dot{x}^2(t) + \dot{y}^2(t)}\, dt = l. \tag{10.43b}$$

This variational problem with the side condition becomes an unconstrained variational problem according to (10.26) with

$$H = H(x, y, \dot{x}, \dot{y}) = y\dot{x} + \lambda\sqrt{\dot{x}^2 + \dot{y}^2}. \tag{10.43c}$$

H satisfies the condition (10.40), so it is a positive homogeneous function of first degree. Furthermore,

$$M = \frac{1}{\dot{y}^2}H_{\dot{x}\dot{x}} = \frac{\lambda}{(\dot{x}^2 + \dot{y}^2)^{3/2}}, \quad H_{\dot{x}y} = 1, \quad H_{x\dot{y}} = 0, \tag{10.43d}$$

holds, so (10.42c) yields the radius of curvature $R = |\lambda|$. Since λ is a constant, the extremals are circles.

10.4 Variational Problems with Functions of Several Variables

10.4.1 Simple Variational Problem

One of the simplest problems with a function of several variables is the following variational problem for a double integral:

$$I[u] = \iint\limits_{(G)} F(x, y, u(x, y), u_x, u_y)\, dx\, dy = \text{extreme!}\,. \tag{10.44}$$

The unknown function $u = u(x, y)$ should take given values on the boundary Γ of the domain G. Analogously to 10.3.2, p. 612, a *comparable function* is to be introduced in the form

$$u(x, y) = u_0(x, y) + \epsilon\eta(x, y), \tag{10.45}$$

where $u_0(x, y)$ is a solution of the variational problem (10.44), which takes the given boundary values, while $\eta(x, y)$ satisfies the condition

$$\eta(x, y) = 0 \quad \text{on the boundary } \Gamma. \tag{10.46}$$

$\eta(x, y)$ together with $u_0(x, y)$ both are differentiable as many times as needed. The quantity ϵ is a parameter.
Next, a surface is to be determined by $u = u(x, y)$, which is close to the solution surface $u_0(x, y)$. $I[u]$ becomes $I(\epsilon)$ with (10.45), i.e., the variational problem (10.44) becomes an extreme value problem which must satisfy the necessary conditions

$$\frac{dI}{d\epsilon} = 0 \qquad \text{for} \quad \epsilon = 0\,. \tag{10.47}$$

From this follows the *Euler differential equation*

$$\frac{\partial F}{\partial u} - \frac{\partial}{\partial x}\left(\frac{\partial F}{\partial u_x}\right) - \frac{\partial}{\partial y}\left(\frac{\partial F}{\partial u_y}\right) = 0 \tag{10.48}$$

as a necessary condition for the solution of the variational problem (10.44).

■ A free membrane, fixed at the perimeter Γ of a domain G in the x, y plane, covers a surface with area

$$I_1 = \iint\limits_{(G)} dx\, dy. \tag{10.49a}$$

If the membrane is deformed by a load so that every point has an elongation $u = u(x, y)$ in the z-direction, then its area is calculated by the formula

$$I_2 = \iint\limits_{(G)} \sqrt{1 + u_x^2 + u_y^2}\, dx\, dy\,. \tag{10.49b}$$

Linearizing the integrand in (10.49b) and using Taylor series (see 6.2.2.3, p. 449), one gets the relation

$$I_2 \approx I_1 + \frac{1}{2}\iint\limits_{(G)} \left(u_x^2 + u_y^2\right)\, dx\, dy\,. \tag{10.49c}$$

For the potential energy U of the deformed membrane holds

$$U = \sigma(I_2 - I_1) = \frac{\sigma}{2}\iint\limits_{(G)} \left(u_x^2 + u_y^2\right)\, dx\, dy\,, \tag{10.49d}$$

where the constant σ denotes the tension of the membrane. In this way arises the so-called *Dirichlet variational problem*: The function $u = u(x, y)$ is to be determined so that the functional

$$I[u] = \iint\limits_{(G)} \left(u_x^2 + u_y^2\right)\, dx\, dy \tag{10.49e}$$

should have an extremum, and u vanishes on the boundary Γ of the plane domain G. The corresponding Euler differential equation is

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0\,. \tag{10.49f}$$

It is the Laplace differential equation for functions of two variables (see 13.5.1, p. 729).

10.4.2 More General Variational Problems

Two generalizations of the simple variational problem are to be considered here.

1. $\boldsymbol{F = F(x, y, u(x,y), u_x, u_y, u_{xx}, u_{xy}, u_{yy})}$

The functional depends on higher-order partial derivatives of the unknown function $u(x,y)$. If the partial derivatives occur up to second order, then the Euler differential equation is:

$$\frac{\partial F}{\partial u} - \frac{\partial}{\partial x}\left(\frac{\partial F}{\partial u_x}\right) - \frac{\partial}{\partial y}\left(\frac{\partial F}{\partial u_y}\right) + \frac{\partial^2}{\partial x^2}\left(\frac{\partial F}{\partial u_{xx}}\right) + \frac{\partial^2}{\partial x \partial y}\left(\frac{\partial F}{\partial u_{xy}}\right) + \frac{\partial^2}{\partial y^2}\left(\frac{\partial F}{\partial u_{yy}}\right) = 0. \tag{10.50}$$

2. $\boldsymbol{F = F(x_1, x_2, \ldots, x_n, u(x_1, \ldots, x_n), u_{x_1}, \ldots, u_{x_n})}$

In the case of a variational problem with n independent variables $x_1, x_2, \ldots, x_n$, the Euler differential equation is:

$$\frac{\partial F}{\partial u} - \sum_{k=1}^{n} \frac{\partial}{\partial x_k}\left(\frac{\partial F}{\partial u_{x_k}}\right) = 0. \tag{10.51}$$

10.5 Numerical Solution of Variational Problems

Most often two ways are used to solve variational problems in practice.

1. Solution of the Euler Differential Equation and Fitting the Found Solution to the Boundary Conditions

Usually, an exact solution of the Euler differential equation is possible only in the simplest cases, so numerical methods are to be used solving the boundary value problem for ordinary or for partial differential equations (see 19.5, p. 976 or 20.3.4, p. 1042ff).

2. Direct Methods

The direct methods start directly from the variational problem and do not use the Euler differential equation. The most popular and probably the oldest procedure is the *Ritz method*. It belongs to the so-called approximation methods which are also used to get approximate solutions of differential equations (see 19.4.2.2, p. 974 and 19.5.2, p. 977). The following example demonstrates this method.

■ Solve numerically the isoperimetric problem

$$\int_0^1 y'^2(x)\,dx = \text{extreme!} \quad (10.52\text{a}) \qquad \text{for} \quad \int_0^1 y^2(x)\,dx = 1 \quad \text{and} \quad y(0) = y(1) = 0. \quad (10.52\text{b})$$

The corresponding variational problem without integral side condition according to 10.3.3, p. 614, is:

$$I[y] = \int_0^1 \left[y'^2(x) - \lambda y^2(x)\right] dx = \text{extreme!} \tag{10.52c}$$

As starting approach to find an approximation solution can be used

$$y(x) = a_1 x(x-1) + a_2 x^2(x-1). \tag{10.52d}$$

Both approximation functions $x(x-1)$ and $x^2(x-1)$ are linearly independent, and satisfy the boundary conditions. Reducing (10.52c) with (10.52d) gives

$$I(a_1, a_2) = \frac{1}{3}a_1^2 + \frac{2}{15}a_2^2 + \frac{1}{3}a_1 a_2 - \lambda\left(\frac{1}{30}a_1^2 + \frac{1}{105}a_2^2 + \frac{1}{30}a_1 a_2\right), \tag{10.52e}$$

and the necessary conditions $\frac{\partial I}{\partial a_1} = \frac{\partial I}{\partial a_2} = 0$ result in he homogeneous linear equation system

$$\left(\frac{2}{3} - \frac{\lambda}{15}\right) a_1 + \left(\frac{1}{3} - \frac{\lambda}{30}\right) a_2 = 0, \qquad \left(\frac{1}{3} - \frac{\lambda}{30}\right) a_1 + \left(\frac{4}{15} - \frac{2\lambda}{105}\right) a_2 = 0. \tag{10.52f}$$

This system has a non-trivial solution only if the determinant of the coefficient matrix is equal to zero:

$$\lambda^2 - 52\lambda + 420 = 0, \quad \text{i.e.,} \quad \lambda_1 = 10,\ \lambda_2 = 42. \tag{10.52g}$$

For $\lambda = \lambda_1 = 10$ from (10.52f) follows $a_2 = 0$, a_1 arbitrary, so the normalized solution belonging to $\lambda_1 = 10$ is:

$$y = 5.48x(x-1)\,. \tag{10.52h}$$

To make a comparison, consider the Euler differential equation belonging to (10.52f). Here the boundary value problem

$$y'' + \lambda y = 0 \quad \text{with} \quad y(0) = y(1) = 0 \tag{10.52i}$$

holds with the eigenvalues $\lambda_k = k^2\pi^2$ $(k = 1, 2, \ldots)$ and the solution $y_k = c_k \sin k\pi x$. The normalized solution, e.g., for the case $k = 1$, i.e., $\lambda_1 = \pi^2 \approx 9.87$ is

$$y = \sqrt{2}\sin \pi x, \tag{10.52j}$$

which is really very close to the approximate solution (10.52h).

Remark: With today's level of computers and science first of all, the *finite element method* (FEM) is to be applied for numerical solutions of variational problems.
The basic idea of this method is given in 19.5.3, p. 978, for numerical solutions of differential equations. The correspondence between differential and variational equations will be used there, e.g., by Euler differential equations or bilinear forms according to (19.146a,b).
Also the *gradient method* can be used for the numerical solution of variational problems as an efficient numerical method for non-linear optimization problems (see 18.2.7, p. 936).

10.6 Supplementary Problems

10.6.1 First and Second Variation

The derivation of the Euler differential equation with a comparable function by the help of the Taylor expansion of the integrand (see 10.3.2, p. 612) was stopped after the linear terms with respect to ϵ:

$$I(\epsilon) = \int_a^b F(x,\ y_0 + \epsilon\eta,\ y_0' + \epsilon\eta')\, dx\,. \tag{10.53}$$

Considering also quadratic terms yields

$$I(\epsilon) - I(0) = \epsilon \int_a^b \left[\frac{\partial F}{\partial y}(x, y_0, y_0')\eta + \frac{\partial F}{\partial y'}(x, y_0, y_0')\eta'\right] dx \tag{10.54}$$

$$+\frac{\epsilon^2}{2}\int_a^b \left[\frac{\partial^2 F}{\partial y^2}(x, y_0, y_0')\eta^2 + 2\frac{\partial^2 F}{\partial y \partial y'}(x, y_0, y_0')\eta\eta' + \frac{\partial^2 F}{\partial y'^2}(x, y_0, y_0')\eta'^2 + O(\epsilon)\right] dx\,.$$

Denoting as

1. Variation δI of the functional $I[y]$ the expression

$$\delta I = \int_a^b \left[\frac{\partial F}{\partial y}(x, y_0, y_0')\eta + \frac{\partial F}{\partial y'}(x, y_0, y_0')\eta'\right] dx \ \text{ and as} \tag{10.55}$$

2. Variation $\delta^2 I$ of the functional $I[y]$ the expression

$$\delta^2 I = \int_a^b \left[\frac{\partial^2 F}{\partial y^2}(x, y_0, y_0')\eta^2 + 2\frac{\partial^2 F}{\partial y \partial y'}(x, y_0, y_0')\eta\eta' + \frac{\partial^2 F}{\partial y'^2}(x, y_0, y_0')\eta'^2 \right] dx, \tag{10.56}$$

then one can write:

$$I(\epsilon) - I(0) \approx \epsilon\, \delta I + \frac{\epsilon^2}{2}\, \delta^2 I. \tag{10.57}$$

By the help of these variations different optimality conditions for the functional $I[y]$ can be formulated (see [10.7]).

10.6.2 Application in Physics

Variational calculus holds its solid place in physics. So, e.g., the fundamental equations of the Newtonian mechanics can be derived from a variational principle and in this way one can come the Hamilton-Jacobi theory. Variational calculus is also important in both atomic theory and quantum physics. It is obvious that the extension and generalization of classical mathematical notions is undoubtedly necessary. Therefore the calculus of variations must be discussed today by modern mathematical disciplines, e.g., functional analysis and optimization. Unfortunately, in this book it is not possible to give more than a brief account of the classical part of the calculus of variations (see [10.4], [10.5], [10.7]).

11 Linear Integral Equations

11.1 Introduction and Classification

1. Definitions

An integral equation is an equation in which the unknown function appears under the integral sign. There is no universal method for solving integral equations. Solution methods and even the existence of a solution depend on the particular form of the integral equation.

An integral equation is called *linear* if linear operations are performed on the unknown function. The *general form* of a linear integral equation is:

$$g(x)\varphi(x) = f(x) + \lambda \int\limits_{a(x)}^{b(x)} K(x,y)\varphi(y)\,dy, \quad c \le x \le d. \tag{11.1}$$

The unknown function is $\varphi(x)$, the function $K(x,y)$ is called the *kernel of the integral equation*, and $f(x)$ is the so-called *perturbation function*. These functions can take complex values as well. The integral equation is *homogeneous* if the function $f(x)$ is identically zero over the considered domain, i.e., $f(x) \equiv 0$, otherwise it is *inhomogeneous*. λ is usually a complex parameter.

Two types of equation (11.1) are of special importance. If the limits of the integral are independent of x, i.e., $a(x) \equiv a$ and $b(x) \equiv b$, it is called a *Fredholm integral equation* (11.2a,11.2b).

If $a(x) \equiv a$ and $b(x) = x$, it is called a Volterra integral equation (11.2c, 11.2d).

If the unknown function $\varphi(x)$ appears only under the integral sign, i.e., $g(x) \equiv 0$ holds, one speaks of an integral equation *of the first kind* as (11.2a), (11.2c). The equation is called an integral equation *of the second kind* if $g(x) \equiv 1$ as in (11.2b), (11.2d).

$$0 = f(x) + \lambda \int\limits_a^b K(x,y)\varphi(y)\,dy, \tag{11.2a}$$

$$\varphi(x) = f(x) + \lambda \int\limits_a^b K(x,y)\varphi(y)\,dy, \tag{11.2b}$$

$$0 = f(x) + \lambda \int\limits_a^x K(x,y)\varphi(y)\,dy, \tag{11.2c}$$

$$\varphi(x) = f(x) + \lambda \int\limits_a^x K(x,y)\varphi(y)\,dy. \tag{11.2d}$$

2. Relations with Differential Equations

The problems of physics and mechanics relative rarely lead directly to an integral equation. These problems can be described mostly by differential equations. The importance of integral equations is that many of these differential equations, together with the initial and boundary values, can be transformed into integral equations.

■ From the initial value problem $y'(x) = f(x,y)$ with $x \ge x_0$ and $y(x_0) = y_0$ by integration from x_0 to x one gets

$$y(x) = y_0 + \int_{x_0}^{x} f(\xi, y(\xi))\,d\xi. \tag{11.3}$$

The unknown function $y(x)$ appears on the left-hand side of (11.3) and also under the integral sign. The integral equation (11.3) is linear if the function $f(\xi, y(\xi))$ has the form $f(\xi, \eta(\xi)) = a(\xi)\,y(\xi) + b(\xi)$, i.e., the original differential equation is also linear.

Remark: This chapter 11 deals with only integral equations of the first and second kind of Fredholm and Volterra types, and also with some singular integral equations.

I.N. Bronshtein et al., *Handbook of Mathematics*,
DOI 10.1007/978-3-662-46221-8_11

11.2 Fredholm Integral Equations of the Second Kind

11.2.1 Integral Equations with Degenerate Kernel

If the *kernel* $K(x, y)$ *of an integral equation* is the finite sum of products of two functions of one variable, i.e., one depends only on x and the other one only on y, it is called a *degenerate kernel* or a *product kernel*.

1. Solution in the Case of a Degenerate Kernel

The solution of a Fredholm integral equation of the second kind with a degenerate kernel leads to the solution of a finite-dimensional system of equations. Consider the integral equation

$$\varphi(x) = f(x) + \lambda \int\limits_a^b K(x, y)\varphi(y)\,dy \quad \text{with} \tag{11.4a}$$

$$K(x, y) = \alpha_1(x)\beta_1(y) + \alpha_2(x)\beta_2(y) + \ldots + \alpha_n(x)\beta_n(y). \tag{11.4b}$$

The functions $\alpha_1(x), \ldots, \alpha_n(x)$ and $\beta_1(x), \ldots, \beta_n(x)$ are given on the interval $[a, b]$ and are supposed to be continuous. Furthermore, the functions $\alpha_1(x), \ldots, \alpha_n(x)$ are supposed to be linearly independent of one another, i.e., the equality

$$\sum_{k=1}^{n} c_k\,\alpha_k(x) \equiv 0 \tag{11.5}$$

with constant coefficients c_k holds for every x in $[a, b]$ only if $c_1 = c_2 = \ldots = c_n = 0$. Otherwise, $K(x, y)$ can be expressed as the sum of a smaller number of products.
From (11.4a) and (11.4b) follows:

$$\varphi(x) = f(x) + \lambda\alpha_1(x) \int\limits_a^b \beta_1(y)\varphi(y)\,dy + \ldots + \lambda\alpha_n(x) \int\limits_a^b \beta_n(y)\varphi(y)\,dy. \tag{11.6a}$$

The integrals are no longer functions of the variable x, they are constant values. Let's denote them by A_k:

$$A_k = \int\limits_a^b \beta_k(y)\varphi(y)\,dy, \quad k = 1, \ldots, n. \tag{11.6b}$$

The solution function $\varphi(x)$, if any exists, is the sum of the perturbation function $f(x)$ and a linear combination of the functions $\alpha_1(x), \ldots, \alpha_n(x)$:

$$\varphi(x) = f(x) + \lambda A_1\alpha_1(x) + \lambda A_2\alpha_2(x) + \ldots + \lambda A_n\alpha_n(x). \tag{11.6c}$$

2. Calculation of the Coefficients of the Solution

The coefficients $A_1, \ldots, A_n$ are calculated as follows. Equation (11.6c) is multiplied by $\beta_k(x)$ and its integral is calculated with respect to x with the limits a and b:

$$\int\limits_a^b \beta_k(x)\varphi(x)\,dx = \int\limits_a^b \beta_k(x)f(x)\,dx + \lambda A_1 \int\limits_a^b \beta_k(x)\alpha_1(x)\,dx + \ldots + \lambda A_n \int\limits_a^b \beta_k(x)\alpha_n(x)\,dx. \tag{11.7a}$$

The left-hand side of this equation is equal to A_k according to (11.6b). Using the following notation

$$b_k = \int\limits_a^b \beta_k(x)f(x)\,dx \quad \text{and} \quad c_{kj} = \int\limits_a^b \beta_k(x)\alpha_j(x)\,dx \tag{11.7b}$$

there is for $k = 1, \ldots, n$:

$$A_k = b_k + \lambda c_{k1}A_1 + \lambda c_{k2}A_2 + \ldots + \lambda c_{kn}A_n. \tag{11.7c}$$

It is possible that the exact values of the integrals in (11.7b) cannot be calculated. When this is the case, their approximate values must be calculated by one of the formulas given in 19.3, p. 963. The linear system of equations (11.7c) contains n equations for the unknown values $A_1, \ldots, A_n$:

$$\begin{array}{llll} (1-\lambda c_{11})A_1 & -\lambda c_{12}\, A_2 - \ldots & & -\lambda c_{1n}\, A_n = b_1\,, \\ -\lambda c_{21}\, A_1 & +(1-\lambda c_{22})A_2 - \ldots & & -\lambda c_{2n}\, A_n = b_2\,, \\ \ldots\ldots\ldots & & & \\ -\lambda c_{n1}\, A_1 & -\lambda c_{n2}\, A_2 - \ldots & +(1-\lambda c_{nn})A_n = b_n\,. & \end{array} \tag{11.7d}$$

3. Analyzing the Solution, Eigenvalues and Eigenfunctions

It is known from the theory of linear systems of equations that (11.7d) has one and only one solution for $A_1, \ldots, A_n$ if the determinant of the matrix of the coefficients is not equal to zero, i.e.,

$$\mathrm{D}(\lambda) = \begin{vmatrix} (1-\lambda c_{11}) & -\lambda c_{12} & \ldots & -\lambda c_{1n} \\ -\lambda c_{21} & (1-\lambda c_{22}) & \ldots & -\lambda c_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ -\lambda c_{n1} & -\lambda c_{n2} & \ldots & (1-\lambda c_{nn}) \end{vmatrix} \neq 0. \tag{11.8}$$

Obviously $\mathrm{D}(\lambda)$ is not identically zero, as $\mathrm{D}(0) = 1$ holds. So there is a number $R > 0$ such that $D(\lambda) \neq 0$ if $|\lambda| < R$. For further investigation two different cases are considered.

Case $\mathbf{D(\lambda) \neq 0}$:

The integral equation has exactly one solution in the form (11.6c), and the coefficients $A_1, \ldots, A_n$ are given by the solution of the system of equations (11.7d). If (11.4a) is a homogeneous integral equation, i.e., $f(x) \equiv 0$, then $b_1 = b_2 = \ldots = b_n = 0$. Then the homogeneous system of equations (11.7d) has only the trivial solution $A_1 = A_2 = \ldots = A_n = 0$. In this case only the function $\varphi(x) \equiv 0$ satisfies the integral equation.

Case $\mathbf{D(\lambda) = 0}$:

$\mathrm{D}(\lambda)$ is a polynomial of no higher than n-th degree, so it can have at most n roots. For these values of λ the homogeneous system of equations (11.7d) with $b_1 = b_2 = \ldots = b_n = 0$ also has non-trivial solutions, so besides the trivial solution $\varphi(x) \equiv 0$ the homogeneous system of equations has other solutions of the form

$$\varphi(x) = C \cdot (A_1\alpha_1(x) + A_2\alpha_2(x) + \ldots + A_n\alpha_n(x)) \quad (C \text{ is an arbitrary constant.})$$

Because $\alpha_1(x), \ldots, \alpha_n(x)$ are linearly independent, $\varphi(x)$ is not identically zero. The roots of $\mathrm{D}(\lambda)$ are called the *eigenvalues* of the integral equation. The corresponding non-vanishing solutions of the homogeneous integral equation are called the *eigenfunctions* belonging to the eigenvalue λ. Several linearly independent eigenfunctions can belong to the same eigenvalue. If the integral equation has a general kernel, then are be to considered all values of λ eigenvalues, for which the homogeneous integral equation has non-trivial solutions. Some authors call the λ with $\mathrm{D}(\lambda) = 0$ the characteristic number, and $\mu = \dfrac{1}{\lambda}$ is called the eigenvalue corresponding to an equation form $\mu\varphi(x) = \int_a^b K(x,y)\varphi(y)\,dy$.

4. Transposed Integral Equation

Now it is necessary to investigate the conditions under which the inhomogeneous integral equation will have solutions if $\mathrm{D}(\lambda) = 0$. For this purpose the *transposed integral equation* (or *adjoint* in the complex case) of (11.4a) is introduced:

$$\psi(x) = g(x) + \lambda \int_a^b K(y,x)\psi(y)\,dy. \tag{11.9a}$$

Let λ be an eigenvalue and $\varphi(x)$ a solution of the inhomogeneous integral equation (11.4a). It is easy to show that λ is also an eigenvalue of the adjoint equation. Now multiply both sides of (11.4a) by any

solution $\psi(x)$ of the homogeneous adjoint integral equation and evaluate the integral with respect to x between the limits a and b:

$$\int_a^b \varphi(x)\psi(x)\,dx = \int_a^b f(x)\psi(x)\,dx + \int_a^b \left(\lambda \int_a^b K(x,y)\psi(x)\,dx\right)\varphi(y)\,dy. \tag{11.9b}$$

Assuming that $\psi(y) = \lambda \int_a^b K(x,y)\psi(x)\,dx$, then $\int_a^b f(x)\psi(x)\,dx = 0$ holds.

That is: The inhomogeneous integral equation (11.4a) has a solution for some eigenvalue λ if and only if the perturbation function $f(x)$ is *orthogonal* to every non-vanishing solution of the homogeneous adjoint integral equation belonging to the same λ. This statement is valid not only for integral equations with degenerate kernels, but also for those with general kernels.

■ **A:** $\varphi(x) = x + \int_{-1}^{+1}(x^2y + xy^2 - xy)\varphi(y)\,dy,\ \alpha_1(x) = x^2,\ \alpha_2(x) = x,\ \alpha_3(x) = -x,\ \beta_1(y) = y,\ \beta_2(y) = y^2,\ \beta_3(y) = y$. The functions $\alpha_k(x)$ are linearly dependent. Therefore one transforms the integral equation into the form $\varphi(x) = x + \int_{-1}^{+1}[x^2y + x(y^2 - y)]\varphi(y)\,dy$. For this integral equation $\alpha_1(x) = x^2,\ \alpha_2(x) = x,\ \beta_1(y) = y,\ \beta_2(y) = y^2 - y$ holds. If any solution $\varphi(x)$ exists, it has the form $\varphi(x) = x + A_1x^2 + A_2\,x$.

$$c_{11} = \int_{-1}^{+1} x^3\,dx = 0, \qquad c_{12} = \int_{-1}^{+1} x^2\,dx = \frac{2}{3}, \qquad b_1 = \int_{-1}^{+1} x^2\,dx = \frac{2}{3},$$
$$c_{21} = \int_{-1}^{+1} (x^4 - x^3)\,dx = \frac{2}{5}, \qquad c_{22} = \int_{-1}^{+1} (x^3 - x^2)\,dx = -\frac{2}{3}, \qquad b_2 = \int_{-1}^{+1} (x^3 - x^2)\,dx = -\frac{2}{3}.$$

With these values the system of equations to determine A_1 and A_2 has the form: $A_1 - \frac{2}{3}A_2 = \frac{2}{3},\ -\frac{2}{5}A_1 + \left(1 + \frac{2}{3}\right)A_2 = -\frac{2}{3}$, which in turn yield that $A_1 = \frac{10}{21},\ A_2 = -\frac{2}{7}$ and $\varphi(x) = x + \frac{10}{21}x^2 - \frac{2}{7}x = \frac{10}{21}x^2 + \frac{5}{7}x$.

■ **B:** $\varphi(x) = x + \lambda \int_0^\pi \sin(x+y)\varphi(y)\,dy$, i.e.: $K(x,y) = \sin(x+y) = \sin x\cos y + \cos x \sin y,\ \varphi(x) = x + \lambda \sin x \int_0^\pi \cos y\,\varphi(y)\,dy + \lambda \cos x \int_0^\pi \sin y\,\varphi(y)\,dy$.

$$c_{11} = \int_0^\pi \sin x \cos x\,dx = 0, \qquad c_{12} = \int_0^\pi \cos^2 x\,dx = \frac{\pi}{2}, \qquad b_1 = \int_0^\pi x\cos x\,dx = -2,$$
$$c_{21} = \int_0^\pi \sin^2 x\,dx = \frac{\pi}{2}, \qquad c_{22} = \int_0^\pi \cos x \sin x\,dx = 0, \qquad b_2 = \int_0^\pi x \sin x\,dx = \pi.$$

With these values the system (11.7d) is $A_1 - \lambda\frac{\pi}{2}A_2 = -2,\ -\lambda\frac{\pi}{2}A_1 + A_2 = \pi$. It has a unique solution for any λ with $D(\lambda) = \begin{vmatrix} 1 & -\lambda\frac{\pi}{2} \\ -\lambda\frac{\pi}{2} & 1 \end{vmatrix} = 1 - \lambda^2\frac{\pi^2}{4} \neq 0$. So $A_1 = \dfrac{\lambda\frac{\pi^2}{2} - 2}{1 - \lambda^2\frac{\pi^2}{4}},\ A_2 = \dfrac{\pi(1-\lambda)}{1 - \lambda^2\frac{\pi^2}{4}}$, and the solution of the integral equation is $\varphi(x) = x + \dfrac{\lambda}{1 - \lambda^2\frac{\pi^2}{4}}\left[\left(\lambda\frac{\pi^2}{2} - 2\right)\sin x + \pi(1-\lambda)\cos x\right]$. The eigenvalues of the integral equation are $\lambda_1 = \frac{2}{\pi},\quad \lambda_2 = -\frac{2}{\pi}$.

The homogeneous integral equation $\varphi(x) = \lambda_k \int_0^\pi \sin(x+y)\varphi(y)\,dy$ has non-trivial solutions of the

form $\varphi_k(x) = \lambda_k(A_1 \sin x + A_2 \cos x)$ $(k = 1, 2)$. For $\lambda_1 = \dfrac{2}{\pi}$ holds $A_1 = A_2$, and with an arbitrary constant A it follows $\varphi_1(x) = A(\sin x + \cos x)$. Similarly for $\lambda_2 = -\dfrac{2}{\pi}$ holds $\varphi_2(x) = B(\sin x - \cos x)$ with an arbitrary constant B.

Remark: This previous solution method is fairly simple but it only works in the case of a degenerate kernel. This method can be used to get a good approximate solution in the case of a general kernel too if it is possible to approximate the general kernel by a degenerate one closely enough.

11.2.2 Successive Approximation Method, Neumann Series

1. Iteration Method

Similarly to the *Picard iteration method* (see 9.1.1.5, **1.**, p. 549) for the solution of ordinary differential equations, an iterative method needs to be given to solve Fredholm integral equations of the second kind. Starting with the equation

$$\varphi(x) = f(x) + \lambda \int_a^b K(x, y)\varphi(y)\, dy, \tag{11.10}$$

one defines a sequence of functions $\varphi_0(x)$, $\varphi_1(x)$, $\varphi_2(x), \ldots$. Let the first be $\varphi_0(x) = f(x)$. The subsequent $\varphi_n(x)$ can be get by the formula

$$\varphi_n(x) = f(x) + \lambda \int_a^b K(x, y)\varphi_{n-1}(y)\, dy \quad (n = 1, 2, \ldots; \quad \varphi_0(x) = f(x)). \tag{11.11a}$$

Following the given method the first step is

$$\varphi_1(x) = f(x) + \lambda \int_a^b K(x, y) f(y)\, dy. \tag{11.11b}$$

According to the iteration formula this expression of $\varphi(y)$ is substituted into the right-hand side of (11.10). To avoid the accidental confusion of the integral variables, y is denoted by η in (11.11b).

$$\varphi_2(x) = f(x) + \lambda \int_a^b K(x, y) \left[f(y) + \lambda \int_a^b K(y, \eta) f(\eta)\, d\eta \right] dy \tag{11.11c}$$

$$= f(x) + \lambda \int_a^b K(x, y) f(y)\, dy + \lambda^2 \int_a^b \int_a^b K(x, y) K(y, \eta) f(\eta)\, dy\, d\eta. \tag{11.11d}$$

Introducing the notation of $K_1(x, y) = K(x, y)$ and $K_2(x, y) = \int_a^b K(x, \xi) K(\xi, y)\, d\xi$, and renaming η as y, one can write $\varphi_2(x)$ in the form

$$\varphi_2(x) = f(x) + \lambda \int_a^b K_1(x, y) f(y)\, dy + \lambda^2 \int_a^b K_2(x, y) f(y)\, dy. \tag{11.11e}$$

Denoting

$$K_n(x, y) = \int_a^b K(x, \xi) K_{n-1}(\xi, y)\, d\xi \quad (n = 2, 3, \ldots) \tag{11.11f}$$

then there is the representation of the n-th iterated $\varphi_n(x)$:

$$\varphi_n(x) = f(x) + \lambda \int_a^b K_1(x,y)f(y)\,dy + \ldots + \lambda^n \int_a^b K_n(x,y)f(y)\,dy. \tag{11.11g}$$

$K_n(x,y)$ is called the *n-th iterated kernel* of $K(x,y)$.

2. Convergence of the Neumann Series

To get the solution $\varphi(x)$, it is to be discussed the convergence of the power series of λ

$$f(x) + \sum_{n=1}^{\infty} \lambda^n \int_a^b K_n(x,y)f(y)\,dy, \tag{11.12}$$

which is called the Neumann series. If the functions $K(x,y)$ and $f(x)$ are bounded, i.e., the inequalities

$$|K(x,y)| < M \quad (a \le x \le b,\ \ a \le y \le b) \ \text{ and } \ |f(x)| < N \quad (a \le x \le b), \tag{11.13a}$$

hold, then the series

$$N \sum_{n=0}^{\infty} |\lambda M(b-a)|^n \tag{11.13b}$$

is a majorant series for the power series (11.12). This geometric series is convergent for all

$$|\lambda| < \frac{1}{M(b-a)}\,. \tag{11.13c}$$

The Neumann series is absolutely and uniformly convergent for all values of λ satisfying (11.13c). By a sharper estimation of the terms of the Neumann series one can give the convergence interval more precisely. According to this, the Neumann series is convergent for

$$|\lambda| < \frac{1}{\sqrt{\int_a^b \int_a^b |K(x,y)|^2\,dx\,dy}}. \tag{11.13d}$$

This restriction for the parameter λ does not mean that there are no solutions for any $|\lambda|$ outside the bounds set by (11.13d), but only that one cannot get it by the Neumann series. The expression

$$\Gamma(x,y;\lambda) = \sum_{n=1}^{\infty} \lambda^{n-1} K_n(x,y) \tag{11.14a}$$

is called the *resolvent* or *solving kernel* of the integral equation. Using the resolvent one gets the solution in the form

$$\varphi(x) = f(x) + \lambda \int_a^b \Gamma(x,y;\lambda) f(y)\,dy. \tag{11.14b}$$

■ For the inhomogeneous Fredholm integral equation of the second kind $\varphi(x) = x + \lambda \int_0^1 xy\,\varphi(y)\,dy$ follows $K_1(x,y) = xy,\ \ K_2(x,y) = \int_0^1 x\eta\,\eta y\,dy = \frac{1}{3}xy,\ \ K_3(x,y) = \frac{1}{9}xy, \ldots,\ \ K_n(x,y) = \frac{xy}{3^{n-1}}$ and from this $\Gamma(x,y;\lambda) = xy\left(\sum_{n=0}^{\infty} \frac{\lambda^n}{3^n}\right)$. With the limit (11.13c) the series is definitely convergent for $|\lambda| < 1$, because $|K(x,y)| \le M = 1$ holds. The resolvent $\Gamma(x,y;\lambda) = \frac{xy}{\left(1 - \frac{\lambda}{3}\right)}$ is a geometric series which is

convergent even for $|\lambda| < 3$. Thus from (11.14b) follows $\varphi(x) = x + \lambda \int_0^1 \frac{xy^2}{\left(1 - \frac{\lambda}{3}\right)} dy = \frac{x}{1 - \frac{\lambda}{3}}$.

Remark: If for a given λ the relation (11.13d) does not hold, then any continuous kernel can be decomposed into the sum of two continuous kernels $K(x, y) = K^1(x, y) + K^2(x, y)$, where $K^1(x, y)$ is a degenerate kernel, and $K^2(x, y)$ is so small that for this kernel (11.13d) holds. In this way one has an exact solution method for any λ which is not an eigenvalue.

11.2.3 Fredholm Solution Method, Fredholm Theorems

11.2.3.1 Fredholm Solution Method

1. Approximate Solution by Discretization

A Fredholm integral equation of the second kind

$$\varphi(x) = f(x) + \lambda \int_a^b K(x, y)\varphi(y)\, dy \tag{11.15}$$

can be approximately represented by a linear system of equations. It should be assumed that the functions $K(x, y)$ and $f(x)$ are continuous for $a \le x \le b$, $a \le y \le b$.
The integral in (11.15) should be approximated by the so-called left-hand rectangular formula (see 19.3.2.1, p. 964). It is also possible to use any other quadrature formula (see 19.3.1, p. 963). An equidistant partition

$$y_k = a + (k-1)h \quad (k = 1, 2, \cdots, n;\ h = \frac{b-a}{n}) \tag{11.16a}$$

yields the approximation

$$\varphi(x) \approx f(x) + \lambda h\, [K(x, y_1)\varphi(y_1) + \ldots + K(x, y_n)\varphi(y_n)]\,. \tag{11.16b}$$

Replacing $\varphi(x)$ in this expression by a function $\overline{\varphi}(x)$ exactly satisfying (11.16b) yields:

$$\overline{\varphi}(x) = f(x) + \lambda h\, [K(x, y_1)\overline{\varphi}(y_1) + \ldots + K(x, y_n)\overline{\varphi}(y_n)]\,. \tag{11.16c}$$

To determine this approximate solution, it is necessary to know the substitution values of $\overline{\varphi}(x)$ at the interpolation nodes $x_k = a + (k-1)h$. Substituting $x = x_1$, $x = x_2, \ldots,$ $x = x_n$ into (11.16c), yields a linear system of equations for the required n substitution values of $\overline{\varphi}(x_k)$. Using the short-hand notations

$$K_{jk} = K(x_j, y_k),\ \varphi_k = \overline{\varphi}(x_k),\ f_k = f(x_k) \tag{11.17a}$$

this system has the form

$$\begin{array}{rrrr} (1-\lambda h K_{11})\varphi_1 & -\lambda h K_{12}\,\varphi_2 - \ldots & -\lambda h K_{1n}\,\varphi_n & = f_1, \\ -\lambda h K_{21}\,\varphi_1 & +(1-\lambda h K_{22})\varphi_2 - \ldots & -\lambda h K_{2n}\,\varphi_n & = f_2, \\ \multicolumn{4}{c}{\dotfill} \\ -\lambda h K_{n1}\,\varphi_1 & -\lambda h K_{n2}\,\varphi_2 - \ldots & +(1-\lambda h K_{nn})\varphi_n & = f_n. \end{array} \tag{11.17b}$$

This system has the determinant of the coefficients

$$\mathrm{D}_n(\lambda) = \begin{vmatrix} (1-\lambda h K_{11}) & -\lambda h K_{12} & \ldots & -\lambda h K_{1n} \\ -\lambda h K_{21} & (1-\lambda h K_{22}) & \ldots & -\lambda h K_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ -\lambda h K_{n1} & -\lambda h K_{n2} & \ldots & (1-\lambda h K_{nn}) \end{vmatrix}. \tag{11.17c}$$

This determinant has the same structure as the determinant of the coefficients in the solution of an integral equation with a degenerate kernel. The system of equations (11.17b) has a unique solution for every λ where $\mathrm{D}_n(\lambda) \ne 0$. The solution gives the approximate substitution values of the unknown function $\varphi\,(x)$ at the interpolation nodes. The values of λ with $\mathrm{D}_n(\lambda) = 0$ are approximations of the

eigenvalues of the integral equations. The solution of (11.17b) can be written in quotient form (see Cramer rule, 4.5.2.3, p. 311):

$$\varphi_k = \frac{\mathrm{D}_n^k(\lambda)}{\mathrm{D}_n(\lambda)} \approx \varphi(x_k), \quad k = 1, \dots, n. \tag{11.18}$$

Here follows $\mathrm{D}_n^k(\lambda)$ from $\mathrm{D}_n(\lambda)$ by replacing the elements of the k-th column by $f_1,\ f_2, \dots,\ f_n$.

2. Calculation of the Resolvent

If n tends to infinity, so does the number of rows and columns of the determinants $\mathrm{D}_n^k(\lambda)$ and $\mathrm{D}_n(\lambda)$, as well. The determinant

$$\mathrm{D}(\lambda) = \lim_{n\to\infty} \mathrm{D}_n(\lambda) \tag{11.19a}$$

is used to get the solution kernel (*resolvent*) $\Gamma(x, y; \lambda)$ (see 11.2.2, p. 625) in the form

$$\Gamma(x, y; \lambda) = \frac{\mathrm{D}(x, y; \lambda)}{\mathrm{D}(\lambda)}. \tag{11.19b}$$

It is true that every root of $\mathrm{D}(\lambda)$ is a pole of $\Gamma(x, y; \lambda)$. Exactly these values of λ, for which $\mathrm{D}(\lambda) = 0$, are the eigenvalues of the integral equation (11.15), and in this case the homogeneous integral equation has non-vanishing solutions, the eigenfunctions belonging to the eigenvalue λ. In the case of $\mathrm{D}(\lambda) \neq 0$, knowing the resolvent $\Gamma(x, y; \lambda)$, an explicit form of the solution is:

$$\varphi(x) = f(x) + \lambda \int_a^b \Gamma(x, y; \lambda) f(y)\, dy = f(x) + \frac{\lambda}{\mathrm{D}(\lambda)} \int_a^b \mathrm{D}(x, y; \lambda) f(y)\, dy. \tag{11.19c}$$

To get the resolvent, one needs the power series of $\mathrm{D}(x, y; \lambda)$ and $\mathrm{D}(\lambda)$ with respect to λ:

$$\Gamma(x, y; \lambda) = \frac{\mathrm{D}(x, y; \lambda)}{\mathrm{D}(\lambda)} = \frac{\sum\limits_{n=0}^{\infty} (-1)^n K_n(x, y) \cdot \lambda^n}{\sum\limits_{n=0}^{\infty} (-1)^n d_n \cdot \lambda^n}, \tag{11.20a}$$

where $d_0 = 1, \quad K_0(x, y) = K(x, y)$. One gets the further coefficients from the recursive formula:

$$d_n = \frac{1}{n} \int_a^b K_{n-1}(x, x)\, dx, \quad K_n(x, y) = K(x, y) \cdot d_n - \int_a^b K(x, t) K_{n-1}(t, y)\, dt. \tag{11.20b}$$

■ **A:** $\varphi(x) = \sin x + \lambda \int_0^{\frac{\pi}{2}} \sin x \cos y\, \varphi(y)\, dy$. The exact solution of this integral equation is $\varphi(x) = \frac{2}{2 - \lambda} \sin x$. For $n = 3$ with $x_1 = 0,\ x_2 = \frac{\pi}{6},\ x_3 = \frac{\pi}{3},\ h = \frac{\pi}{6}$ gives

$$\mathrm{D}_3(\lambda) = \begin{vmatrix} 1 & 0 & 0 \\ -\frac{\lambda\pi}{12} & 1 - \frac{\sqrt{3}\lambda\pi}{24} & -\frac{\lambda\pi}{24} \\ -\frac{\sqrt{3}\lambda\pi}{12} & -\frac{3\lambda\pi}{24} & 1 - \frac{\sqrt{3}\lambda\pi}{24} \end{vmatrix} = \left(1 - \frac{\sqrt{3}\lambda\pi}{24}\right)^2 - \frac{\lambda^2\pi^2}{192} = 1 - \frac{\sqrt{3}\lambda\pi}{12}.\ \lambda = \frac{12}{\sqrt{3}\pi} \approx 2.205$$

is an approximation of the exact eigenvalue $\lambda = 2$. From the first equation of the system of equations (11.17b) for $f_1 = 0$ follows the solution $\varphi_1 = 0$. Substituting this result into the second and third equation gives the system of equations: $\left(1 - \frac{\sqrt{3}\lambda\pi}{24}\right)\varphi_2 - \frac{\lambda\pi}{24}\varphi_3 = \frac{1}{2}, \quad -\frac{3\lambda\pi}{24}\varphi_2 + \left(1 - \frac{\sqrt{3}\lambda\pi}{24}\right)\varphi_3 = \frac{\sqrt{3}}{2}$. This system has the solution $\varphi_2 = \frac{1}{2 - \frac{\sqrt{3}\pi}{6}\lambda}$, $\varphi_3 = \frac{\sqrt{3}}{2 - \frac{\sqrt{3}\pi}{6}\lambda}$. If $\lambda = 1$, then $\varphi_1 = 0, \quad \varphi_2 =$

0.915, $\varphi_3 = 1.585$. The substitution values of the exact solution are: $\varphi(0) = 0$, $\varphi\left(\frac{\pi}{6}\right) = 1$, $\varphi\left(\frac{\pi}{3}\right) =$ 1.732.

In order to achieve better accuracy, the number of interpolation nodes needs to be increased.

■ **B:** $\varphi(x) = x + \lambda \int_0^1 (4xy - x^2)\varphi(y)\,dy$; $d_0 = 1$, $K_0(x,y) = 4xy - x^2$, $d_1 = \int_0^1 3x^2\,dx = 1$, $K_1(x,y) = 4xy - x^2 - \int_0^1 (4xt - x^2)(4ty - t^2)\,dt = x + 2x^2y - \frac{4}{3}x^2 - \frac{4}{3}xy$, $d_2 = \frac{1}{2}\int_0^1 K_1(x,x)\,dx = \frac{1}{18}$, $K_2(x,y) = \frac{1}{18}(4xy - x^2) - \int_0^1 K(x,t)K_1(t,y)\,dt = 0$. With these the values d_3, $K_3(x,y)$ and all the following values of d_k and $K_k(x,y)$ are equal to zero. $\Gamma(x,y;\lambda) = \dfrac{4xy - x^2 - \left[x + 2x^2y - \frac{4}{3}x^2 - \frac{4}{3}xy\right]\lambda}{1 - \lambda + \frac{\lambda^2}{18}}$.

From $1 - \lambda + \frac{\lambda^2}{18} = 0$ the two eigenvalues $\lambda_{1,2} = 9 \pm 3\sqrt{7}$ follow. If λ is not an eigenvalue, then the solution is $\varphi(x) = x + \lambda \int_0^1 \Gamma(x,y;\lambda) f(y)\,dy = \dfrac{3x(2\lambda - 3\lambda x + 6)}{\lambda^2 - 18\lambda + 18}$.

11.2.3.2 Fredholm Theorems

For the Fredholm integral equation of the second kind

$$\varphi(x) = f(x) + \lambda \int_a^b K(x,y)\varphi(y)\,dy \tag{11.21a}$$

the correspondent transposed integral equation is given by

$$\psi(x) = g(x) + \lambda \int_a^b K(y,x)\psi(y)\,dy. \tag{11.21b}$$

For this pair of integral equations the following statements are valid (see also 11.2.1, p. 622).

1. A Fredholm integral equation of the second kind can only have finite or countably infinite eigenvalues. The eigenvalues cannot accumulate in any finite interval, i.e., for any positive R there are only a finite number of λ for which $|\lambda| < R$.

2. If λ is not an eigenvalue of (11.21a), then both of the inhomogeneous integral equations have a unique solution for any perturbation function $f(x)$ or $g(x)$, and the corresponding homogeneous integral equations have only trivial solutions.

3. If λ is an eigenvalue of (11.21a), then λ is also an eigenvalue of the transposed equation (11.21b). Both homogeneous integral equations have non-vanishing solutions, and the number of linearly independent solutions are the same for both equations.

4. For an eigenvalue λ the homogeneous integral equation can be solved if and only if the perturbation function is orthogonal to every solution of the homogeneous transposed integral equation, i.e., for every solution of the integral equation

$$\psi(x) = \lambda \int_a^b K(y,x)\psi(y)\,dy, \tag{11.22a}$$

$$\int_a^b f(x)\psi(x)\,dx = 0 \quad \text{holds.} \tag{11.22b}$$

The *Fredholm alternative theorem* follows from these statements: Either the inhomogeneous integral equation can be solved for any perturbation function $f(x)$ or the corresponding homogeneous equation

has non-trivial solutions.

11.2.4 Numerical Methods for Fredholm Integral Equations of the Second Kind

Often it is either impossible or takes too much work to get the exact solution of a Fredholm integral equation of the second kind

$$\varphi(x) = f(x) + \lambda \int_a^b K(x,y)\varphi(y)\,dy \tag{11.23}$$

by the solution methods given in 11.2.1, p. 622, 11.2.2, p. 625 and 11.2.3, p. 627. In such cases certain numerical methods can be used for approximation. Three different methods are given below to get the numerical solution of an integral equation of the form (11.23).

11.2.4.1 Approximation of the Integral

1. Semi-Discrete Problem

Working on the integral equation (11.23) often one replaces the integral by an approximation formula. These approximation formulas are called *quadrature formulas*. They take the form

$$\int_a^b f(x)\,dx \approx Q_{[a,b]}(f) = \sum_{k=1}^{n} \omega_k f(x_k), \tag{11.24}$$

i.e., instead of the integral there is now a sum of the substitution values of the function at the *interpolation nodes* x_k weighted by the values ω_k. The numbers ω_k should be suitably chosen (so as to be independent of f). Equation (11.23) can be written in the approximate form:

$$\varphi(x) \approx f(x) + \lambda Q_{[a,b]}(K(x,\cdot)\varphi(\cdot)) = f(x) + \lambda \sum_{k=1}^{n} \omega_k K(x,y_k)\varphi(y_k). \tag{11.25a}$$

The quadrature formula $Q_{[a,b]}(K(x,\cdot)\varphi(\cdot))$ also depends on the variable x. The dot in the argument of the function means that the quadrature formula will be used with respect to the variable y. Defining the relation

$$\overline{\varphi}(x) = f(x) + \lambda \sum_{k=1}^{n} \omega_k K(x,y_k)\overline{\varphi}(y_k). \tag{11.25b}$$

$\overline{\varphi}(x)$ is an approximation of the exact solution $\varphi(x)$. One considers (11.25b) as a *semi-discrete problem*, because the variable y is turned into discrete values while the variable x can still be arbitrary.

If the equation (11.25b) holds for a function $\overline{\varphi}(x)$ for every $x \in [a,b]$, it must also be valid for the interpolation nodes $x = x_k$:

$$\overline{\varphi}(x_k) = f(x_k) + \lambda \sum_{j=1}^{n} \omega_j K(x_k,y_j)\overline{\varphi}(y_j), \quad k = 1,2,\ldots,n. \tag{11.25c}$$

This is a linear system of equations containing n equations for the n unknown values $\overline{\varphi}(x_k)$. Substituting these solutions into (11.25b) yields the solution of the semi-discrete problem. The accuracy and the amount of calculations of this method depend on the quadrature formula used. For example using the *left-hand rectangular formula* (see 19.3.2.1, p. 964) with an equidistant partition $y_k = x_k = a + h(k-1),\ h = (b-a)/n,\ (k = 1,\ldots,n)$ yields:

$$\int_a^b K(x,y)\overline{\varphi}(y)dy \approx \sum_{k=1}^{n} hK(x,y_k)\overline{\varphi}(y_k). \tag{11.26a}$$

With the notations

$$K_{jk} = K(x_j,y_k), \quad f_k = f(x_k), \quad \varphi_k = \overline{\varphi}(x_k) \tag{11.26b}$$

the system (11.25c) has the form:

$$\begin{array}{llll}(1-\lambda hK_{11})\varphi_1 & -\lambda hK_{12}\,\varphi_2 - \;\ldots & -\lambda hK_{1n}\,\varphi_n & = f_1,\\ -\lambda hK_{21}\,\varphi_1 & +(1-\lambda hK_{22}\varphi_2) - \;\ldots & -\lambda hK_{2n}\,\varphi_n & = f_2,\\ \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\\ -\lambda hK_{n1}\,\varphi_1 & -\lambda hK_{n2}\,\varphi_2 - \;\ldots & +(1-\lambda hK_{nn})\varphi_n & = f_n.\end{array} \tag{11.26c}$$

The same system was involved in the Fredholm solution method (see 11.2.3, p. 627). As the rectangular formula is not accurate enough, for a better approximation of the integral one can increase the number of interpolation nodes, along with an increase in the dimension of the system of equations. Hence one gets the idea of looking for another quadrature formula.

2. Nyström Method

In the so-called *Nyström method* the Gauss quadrature formula is to be used for the approximation of the integral (see 19.3.3, p. 965). In order to derive this, one considers the integral

$$I = \int_a^b f(x)\,dx. \tag{11.27a}$$

The integrand is replaced by a polynomial $p(x)$, namely the interpolation polynomial of $f(x)$ at the interpolation nodes x_k:

$$p(x) = \sum_{k=1}^{n} L_k(x) f(x_k) \text{ with}$$

$$L_k(x) = \frac{(x-x_1)\,\ldots\,(x-x_{k-1})(x-x_{k+1})\,\ldots\,(x-x_n)}{(x_k-x_1)\ldots(x_k-x_{k-1})(x_k-x_{k+1})\ldots(x_k-x_n)}. \tag{11.27b}$$

For this polynomial

$$p(x_k) = f(x_k), \quad k = 1,\ldots,n. \tag{11.27c}$$

holds. The replacement of the integrand $f(x)$ by $p(x)$ results in the quadrature formula

$$\int_a^b f(x)\,dx \approx \int_a^b p(x)\,dx = \sum_{k=1}^{n} f(x_k)\int_a^b L_k(x)\,dx = \sum_{k=1}^{n}\omega_k f(x_k) \text{ with } \omega_k = \int_a^b L_k(x)\,dx. \tag{11.27d}$$

For the Gauss quadrature formula the interpolation nodes cannot be chosen arbitrarily but they must be chosen by the formula:

$$x_k = \frac{a+b}{2} + \frac{b-a}{2}t_k, \quad k = 1,2,\ldots,n. \tag{11.28a}$$

The n values t_k are the n roots of the Legendre polynomial of the first kind (see 9.1.2.6, **3.**, p. 565)

$$P_n(t) = \frac{1}{2^n\cdot n!}\,\frac{d^n\left[(t^2-1)^n\right]}{dt^n}. \tag{11.28b}$$

These roots are in the interval $[-1,+1]$. The coefficients ω_k can be calculated by the substitution $x - x_k = \dfrac{b-a}{2}(t-t_k)$, so:

$$\omega_k = \int_a^b L_k(x)\,dx = (b-a)\,\frac{1}{2}\int_{-1}^{1}\frac{(t-t_1)\ldots(t-t_{k-1})(t-t_{k+1})\ldots(t-t_n)}{(t_k-t_1)\ldots(t_k-t_{k-1})(t_k-t_{k+1})\ldots(t_k-t_n)}\,dt$$
$$= (b-a)A_k. \tag{11.29}$$

In **Table 11.1** are given the roots of the Legendre polynomial of the first kind and the weights A_k for $n = 1,\ldots,6$.

Table 11.1 Roots of the Legendre polynomial of the first kind

n	t	A	n	t	A
1	$t_1 = 0$	$A_1 = 1$	5	$t_1 = -0.9062$	$A_1 = 0.1185$
2	$t_1 = -0.5774$	$A_1 = 0.5$		$t_2 = -0.5384$	$A_2 = 0.2393$
	$t_2 = 0.5774$	$A_2 = 0.5$		$t_3 = 0$	$A_3 = 0.2844$
3	$t_1 = -0.7746$	$A_1 = 0.2778$		$t_4 = 0.5384$	$A_4 = 0.2393$
	$t_2 = 0$	$A_2 = 0.4444$		$t_5 = 0.9062$	$A_5 = 0.1185$
	$t_3 = 0.7746$	$A_3 = 0.2778$	6	$t_1 = -0.9324$	$A_1 = 0.0857$
4	$t_1 = -0.8612$	$A_1 = 0.1739$		$t_2 = -0.6612$	$A_2 = 0.1804$
	$t_2 = -0.3400$	$A_2 = 0.3261$		$t_3 = -0.2386$	$A_3 = 0.2340$
	$t_3 = 0.3400$	$A_3 = 0.3261$		$t_4 = 0.2386$	$A_4 = 0.2340$
	$t_4 = 0.8612$	$A_4 = 0.1739$		$t_5 = 0.6612$	$A_5 = 0.1804$
				$t_6 = 0.9324$	$A_6 = 0.0857$

■ Solve the integral equation $\varphi(x) = \cos \pi x + \dfrac{x}{x^2+\pi^2}(\mathrm{e}^x + 1) + \int_0^1 e^{xy}\varphi(y)\,dy$ by the Nyström method for $n = 3$.

$$\begin{array}{llll} n=3: & x_1 = 0.1127, & x_2 = 0.5, & x_3 = 0.8873, \\ & A_1 = 0.2778, & A_2 = 0.4444, & A_3 = 0.2778, \\ & f_1 = 0.96214, & f_2 = 0.13087, & f_3 = -0.65251, \\ & K_{11} = 1.01278, & K_{22} = 1.28403, & K_{33} = 2.19746, \\ & \multicolumn{3}{l}{K_{12} = K_{21} = 1,05797,\ K_{13} = K_{31} = 1.10517,\ K_{23} = K_{32} = 1.55838.} \end{array}$$

The system of equations (11.25c) for $\varphi_1,\ \varphi_2$, and φ_3 is

$$\begin{aligned} 0.71864\varphi_1 - 0.47016\varphi_2 - 0.30702\varphi_3 &= 0.96214, \\ -0.29390\varphi_1 + 0.42938\varphi_2 - 0.43292\varphi_3 &= 0.13087, \\ -0.30702\varphi_1 - 0.69254\varphi_2 + 0.38955\varphi_3 &= -0.65251. \end{aligned}$$

The solution of the system is: $\varphi_1 = 0.93651,\ \ \varphi_2 = -0.00144,\ \ \varphi_3 = -0.93950$. The substitution values of the exact solution at the interpolation nodes are: $\varphi(x_1) = 0.93797,\ \ \varphi(x_2) = 0,\ \ \varphi(x_3) = -0.93797$.

11.2.4.2 Kernel Approximation

Replace the kernel $K(x, y)$ by a kernel $\overline{K}(x, y)$ so that $\overline{K}(x, y) \approx K(x, y)$ for $a \le x \le b$, $a \le y \le b$. Try to choose a kernel making the solution of the integral equation

$$\overline{\varphi}(x) = f(x) + \lambda \int_a^b \overline{K}(x, y)\overline{\varphi}(y)\,dy \tag{11.30}$$

the easiest possible.

1. Tensor Product Approximation

A frequently-used approximation of the kernel is the *tensor product approximation* in the form

$$K(x, y) \approx \overline{K}(x, y) = \sum_{j=0}^{n}\sum_{k=0}^{n} d_{jk}\,\alpha_j(x)\beta_k(y) \tag{11.31a}$$

with given linearly independent functions $\alpha_0(x), \ldots, \alpha_n(x)$ and $\beta_0(y), \ldots, \beta_n(y)$ whose coefficients d_{jk} must be chosen so that the double sum approximates the kernel closely enough in a certain sense.

Rewriting (11.31a) in a degenerate kernel gives:

$$\overline{K}(x,y) = \sum_{j=0}^{n} \alpha_j(x) \left[\sum_{k=0}^{n} d_{jk}\beta_k(y)\right], \; \delta_j(y) = \sum_{k=0}^{n} d_{jk}\beta_k(y) \; \overline{K}(x,y) = \sum_{j=0}^{n} \alpha_j(x)\delta_j(y). \tag{11.31b}$$

Now, the solution method 11.2.1, p. 622 can be used for the integral equation

$$\overline{\varphi}(x) = f(x) + \lambda \int_a^b \left[\sum_{j=0}^{n} \alpha_j(x)\delta_j(y)\right] \overline{\varphi}(y)\, dy. \tag{11.31c}$$

Functions $\alpha_0(x), \ldots \alpha_n(x)$ and $\beta_0(y), \ldots, \beta_n(y)$ should be chosen so that the coefficients d_{jk} in (11.31a) can be calculated easily and also that the solution of (11.31c) isn't too difficult.

2. Special Spline Approach

One chooses

$$\alpha_k(x) = \beta_k(x) = \begin{cases} 1 - n\left|x - \dfrac{k}{n}\right| & \text{for} \quad \dfrac{k-1}{n} \le x \le \dfrac{k+1}{n}, \\ 0 & \text{otherwise} \end{cases} \tag{11.32}$$

for a special kernel approximation on the interval of integration $[a, b] = [0, 1]$. The function $\alpha_k(x)$ has non-zero values only in the so called *carrier interval* $\left(\dfrac{k-1}{n}, \dfrac{k+1}{n}\right)$, (**Fig. 11.1**).

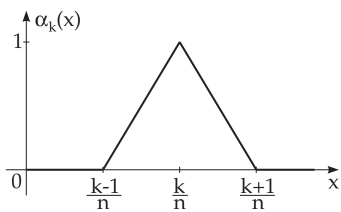


Figure 11.1

To calculate the coefficients d_{jk} in (11.31a), consider $\overline{K}(x, y)$ at the points $x = l/n$, $y = i/n$ $(l, i = 0, 1, \ldots, n)$. Then

$$\alpha_j\left(\frac{l}{n}\right)\alpha_k\left(\frac{i}{n}\right) = \begin{cases} 1 \text{ for } j = l,\; k = i, \\ 0 \;\text{ otherwise} \end{cases} \tag{11.33}$$

holds, and consequently $\overline{K}(l/n, i/n) = d_{li}$. Hence, one substitutes $d_{li} = \overline{K}\left(\frac{l}{n}, \frac{i}{n}\right) = K\left(\frac{l}{n}, \frac{i}{n}\right)$. Now (11.31a) has the form

$$\overline{K}(x,y) = \sum_{j=0}^{n}\sum_{k=0}^{n} K\left(\frac{j}{n}, \frac{k}{n}\right)\alpha_j(x)\beta_k(y). \tag{11.34}$$

As known, the solution of (11.31c) has the representation

$$\overline{\varphi}(x) = f(x) + A_0\alpha_0(x) + \ldots + A_n\alpha_n(x). \tag{11.35}$$

The expression $A_0\alpha_0(x) + \ldots + A_n\alpha_n(x)$ is a piecewise linear function with substitution values A_k at the points $x_k = k/n$. Solving (11.31c) by the method given for the degenerate kernel, gives a linear system of equations for the numbers $A_0, \ldots, A_n$:

$$\begin{array}{llll} (1 - \lambda c_{00})A_0 & -\lambda c_{01}\, A_1 - \;\ldots & -\lambda c_{0n}\, A_n & = b_0, \\ -\lambda c_{10}\, A_0 & +(1 - \lambda c_{11}A_1) - \;\ldots & -\lambda c_{1n}\, A_n & = b_1, \\ \ldots\ldots\ldots\ldots & \ldots\ldots\ldots\ldots & \ldots\ldots\ldots\ldots & \\ -\lambda c_{n0}\, A_0 & -\lambda c_{n1}\, A_1 - \;\ldots & +(1 - \lambda c_{nn})A_n & = b_n, \end{array} \tag{11.36a}$$

where

$$c_{jk} = \int_0^1 \delta_j(x)\alpha_k(x)\, dx = \int_0^1 \left[\sum_{i=0}^{n} K\left(\frac{j}{n}, \frac{i}{n}\right)\alpha_j(x)\right]\alpha_k(x)\, dx$$

$$= K\left(\frac{j}{n}, \frac{0}{n}\right) \int\limits_0^1 \alpha_0(x)\alpha_k(x)\,dx + \ldots + K\left(\frac{j}{n}, \frac{n}{n}\right) \int\limits_0^1 \alpha_n(x)\alpha_k(x)\,dx. \tag{11.36b}$$

For the integrals it holds

$$I_{jk} = \int\limits_0^1 \alpha_j(x)\alpha_k(x)\,dx = \begin{cases} \frac{1}{3n} & \text{for} \quad j=0,\ k=0 \text{ and } j=n,\ k=n, \\ \frac{2}{3n} & \text{for} \quad j=k,\ 1 \le j < n, \\ \frac{1}{6n} & \text{for} \quad j=k+1,\ j=k-1, \\ 0 & \text{otherwise.} \end{cases} \tag{11.36c}$$

The numbers b_k in (11.36a) are given by

$$b_k = \int\limits_0^1 f(x) \left[\sum_{j=0}^{n} K\left(\frac{k}{n}, \frac{j}{n}\right) \alpha_j(x)\right] dx. \tag{11.36d}$$

Taking a matrix $\mathbf{C}$ with numbers c_{jk} from (11.36a), a matrix $\mathbf{B}$ with the values $K(j/n, k/n)$ and a matrix $\mathbf{A}$ with the values I_{jk} respectively, a vector $\underline{\mathbf{b}}$ from the numbers $b_0, \ldots, b_n$, and a vector $\underline{\mathbf{a}}$ from the unknown values $A_0, \ldots, A_n$, the system of equations (11.36a) has the form

$$(\mathbf{I} - \lambda\mathbf{C})\underline{\mathbf{a}} = (\mathbf{I} - \lambda\mathbf{BA})\underline{\mathbf{a}} = \underline{\mathbf{b}}. \tag{11.36e}$$

In the case when the matrix $(\mathbf{I} - \lambda\mathbf{BA})$ is regular, this system has a unique solution $\underline{\mathbf{a}} = (A_0, \ldots, A_n)$.

11.2.4.3 Collocation Method

Suppose the n functions $\varphi_1(x), \ldots, \varphi_n(x)$ are linearly independent in the interval $[a, b]$. They can be used to form an approximation function $\overline{\varphi}(x)$ of the solution $\varphi(x)$:

$$\varphi(x) \approx \overline{\varphi}(x) = a_1\varphi_1(x) + a_2\varphi_2(x) + \ldots + a_n\varphi_n(x). \tag{11.37a}$$

The problem is now to determine the coefficients $a_1, \ldots, a_n$. Usually, there are no values $a_1, \ldots, a_n$ such that the function $\overline{\varphi}(x)$ given in this form represents the exact solution $\varphi(x) = \overline{\varphi}(x)$ of the integral equation (11.23). Therefore, n interpolation points $x_1, \ldots, x_n$ are defined in the interval of integration, and it is required that the approximation function (11.37a) satisfies the integral equation at least at these points:

$$\overline{\varphi}(x_k) = a_1\varphi_1(x_k) + \ldots + a_n\varphi_n(x_k) \tag{11.37b}$$

$$= f(x_k) + \lambda \int\limits_a^b K(x_k, y)\left[a_1\varphi_1(y) + \ldots + a_n\varphi_n(y)\right] dy \quad (k = 1, \ldots, n). \tag{11.37c}$$

With some transformations this system of equations takes the form:

$$\left[\varphi_1(x_k) - \lambda \int\limits_a^b K(x_k, y)\varphi_1(y)\,dy\right] a_1 + \ldots + \left[\varphi_n(x_k) - \lambda \int\limits_a^b K(x_k, y)\varphi_n(y)\,dy\right] a_n$$

$$= f(x_k) \quad (k = 1, \ldots, n). \tag{11.37d}$$

Defining the matrices

$$\mathbf{A} = \begin{pmatrix} \varphi_1(x_1) & \cdots & \varphi_n(x_1) \\ \vdots & & \vdots \\ \varphi_1(x_n) & \cdots & \varphi_n(x_n) \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} \beta_{11} & \cdots & \beta_{1n} \\ \vdots & & \vdots \\ \beta_{n1} & \cdots & \beta_{nn} \end{pmatrix} \text{ with } \beta_{jk} = \int\limits_a^b K(x_j, y)\varphi_k(y)\,dy \tag{11.37e}$$

and the vectors

$$\underline{\mathbf{a}} = (a_1, \ldots, a_n)^\top, \quad \underline{\mathbf{b}} = (f(x_1), \ldots, f(x_n))^\top. \tag{11.37f}$$

then the system of equations to determine the numbers $a_1, \ldots, a_n$ can be written in matrix form:

$$(\mathbf{A} - \lambda\mathbf{B})\ \underline{\mathbf{a}} = \underline{\mathbf{b}}. \tag{11.37g}$$

■ $\varphi(x) = \dfrac{\sqrt{x}}{2} + \displaystyle\int_0^1 \sqrt{xy}\,\varphi(y)\,dy$. The approximation function is $\overline{\varphi}(x) = a_1x^2 + a_2x + a_3,\ \varphi_1(x) = x^2,\ \varphi_2(x) = x,\ \varphi_3(x) = 1$. The interpolation nodes are $x_1 = 0, x_2 = 0.5, x_3 = 1$.

$$\mathbf{A} = \begin{pmatrix} 0 & 0 & 1 \\ \frac{1}{4} & \frac{1}{2} & 1 \\ 1 & 1 & 1 \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} 0 & 0 & 0 \\ \frac{\sqrt{2}}{7} & \frac{\sqrt{2}}{5} & \frac{\sqrt{2}}{3} \\ \frac{2}{7} & \frac{2}{5} & \frac{2}{3} \end{pmatrix}, \quad \underline{\mathbf{b}} = \begin{pmatrix} 0 \\ \frac{1}{2\sqrt{2}} \\ \frac{1}{2} \end{pmatrix}.$$

The system of equations is

$$\begin{aligned} a_3 &= 0, \\ \left(\frac{1}{4} - \frac{\sqrt{2}}{7}\right) a_1 + \left(\frac{1}{2} - \frac{\sqrt{2}}{5}\right) a_2 + \left(1 - \frac{\sqrt{2}}{3}\right) a_3 &= \frac{1}{2\sqrt{2}}, \\ \frac{5}{7}\, a_1 + \frac{3}{5}\, a_2 + \frac{1}{3}\, a_3 &= \frac{1}{2}, \end{aligned}$$

whose solutions are $a_1 = -0.8197, a_2 = 1.8092, a_3 = 0$ and with these $\overline{\varphi}(x) = -0.8197\,x^2 + 1.8092\,x$, and so $\overline{\varphi}(0) = 0,\ \overline{\varphi}(0.5) = 0.6997,\ \overline{\varphi}(1) = 0.9895$.
The exact solution of the integral equation is $\varphi(x) = \sqrt{x}$ with the values $\varphi(0) = 0,\ \varphi(0.5) = 0.7071,\ \varphi(1) = 1$.
In order to improve the accuracy in this example, it is not a good idea to increase the degree of the polynomial, as polynomials of higher degree are numerically unstable. It is much better to use different spline approximations, e.g., a piecewise linear approximation $\overline{\varphi}(x) = a_1\varphi_1(x) + a_2\varphi_2(x) + \cdots + a_n\varphi_n(x)$ with the functions introduced in 11.2.4.2

$$\varphi_k(x) = \begin{cases} 1 - n\left|x - \dfrac{k}{n}\right| & \text{for } \dfrac{k-1}{n} \le x \le \dfrac{k+1}{n}, \\ 0 & \text{otherwise.} \end{cases}$$

In this case, the solution $\varphi(x)$ is approximated by a polygon $\overline{\varphi}(x)$.

Remark: There is no theoretical restriction as to the choice of the interpolation nodes for the collocation method. In the case, however, when the solution function oscillates considerably in a subinterval the number of interpolation points in this interval should be increased.

11.3 Fredholm Integral Equations of the First Kind

11.3.1 Integral Equations with Degenerate Kernels

1. Formulation of the Problem

Consider the Fredholm integral equation of the first kind with degenerate kernel

$$f(x) = \int_a^b (\alpha_1(x)\beta_1(y) + \ldots + \alpha_n(x)\beta_n(y))\varphi(y)\,dy \quad (c \le x \le d), \tag{11.38a}$$

and introduce the notation similar to that used in 11.2, p. 622,

$$A_j = \int_a^b \beta_j(y)\varphi(y)\,dy \quad (j = 1, 2, \ldots, n). \tag{11.38b}$$

Then (11.38a) has the form

$$f(x) = A_1\alpha_1(x) + \ldots + A_n\alpha_n(x), \tag{11.38c}$$

i.e., the integral equation has a solution only if $f(x)$ is a linear combination of the functions $\alpha_1(x), \ldots, \alpha_n(x)$. If this assumption is fulfilled, the constants $A_1, \ldots, A_n$ are known.

2. Initial Approach

Looking for the solution in the form

$$\varphi(x) = c_1\beta_1(x) + \ldots + c_n\beta_n(x) \tag{11.39a}$$

where the coefficients $c_1, \ldots, c_n$ are unknown, substituting in (11.38b)

$$A_i = c_1 \int_a^b \beta_i(y)\beta_1(y)\,dy + \ldots + c_n \int_a^b \beta_i(y)\beta_n(y)\,dy \quad (i = 1, 2, \ldots, n), \tag{11.39b}$$

and introducing the notation

$$K_{ij} = \int_a^b \beta_i(y)\beta_j(y)\,dy \tag{11.39c}$$

gives the following system of equations for the unknown coefficients $c_1, \ldots, c_n$:

$$\begin{aligned} K_{11}c_1 + \ldots + K_{1n}c_n &= A_1, \\ \vdots \qquad\qquad \vdots \quad &\quad \vdots \\ K_{n1}c_1 + \ldots + K_{nn}c_n &= A_n. \end{aligned} \tag{11.39d}$$

3. Solutions

The matrix of the coefficients is non-singular if the functions $\beta_1(y), \ldots, \beta_n(y)$ are linearly independent (see 12.1.3, p. 656). However, the solution obtained in (11.39a) is not the only one. Unlike the integral equations of the second kind with a degenerate kernel, the homogeneous integral equation belonging to (11.38a) always has a non-trivial solution. Suppose $\varphi^h(x)$ is such a solution of the homogeneous equation and $\varphi(x)$ is a solution of (11.38a). Then $\varphi(x) + \varphi^h(x)$ is also a solution of (11.38a).

To determine all the solutions of the homogeneous equation, consider the equation (11.38c) with $f(x) = 0$. If the functions $\alpha_1(x), \ldots, \alpha_n(x)$ are linearly independent, the equation holds if and only if

$$A_j = \int_a^b \beta_j(y)\varphi(y)\,dy = 0 \quad (j = 1, 2, \ldots, n), \tag{11.40}$$

i.e., every function $\varphi^h(y)$ orthogonal to every function $\beta_j(y)$ is a solution of the homogeneous integral equation.

11.3.2 Analytic Basis

1. Initial Approach

Several methods for the solution of Fredholm integral equations of the first kind

$$f(x) = \int_a^b K(x, y)\varphi(y)\,dy \quad (c \le x \le d) \tag{11.41}$$

determine the solution $\varphi(y)$ as a function series of a given system of functions $(\beta_n(y)) = \{\beta_1(y), \beta_2(y), \ldots\}$, i.e., looking for the solution in the form

$$\varphi(y) = \sum_{j=1}^{\infty} c_j\beta_j(y) \tag{11.42}$$

where there are to determine the unknown constants c_j. Choosing the system of functions it is to be considered that the functions $(\beta_n(y))$ should generate the whole space of solutions, and also that the

calculation of the coefficients c_j should be easy.
For an easier survey only real functions are discussed in this section. All of the statements can be extended to complex-valued functions, too. Because of the solution method there are to establish certain requirements for properties of the kernel function $K(x,y)$ (see [11.3], [11.11], [11.12]). It is to be assumed that these requirements are always fulfilled. Next, there are to discuss some relevant information.

2. Quadratically Integrable Functions

A function $\psi(y)$ is *quadratically integrable* over the interval $[a,b]$ if

$$\int_a^b |\psi(y)|^2\,dy < \infty \tag{11.43}$$

holds. For example, every continuous function on $[a,b]$ is quadratically integrable. The space of quadratically integrable functions over $[a,b]$ will be denoted by $L^2[a,b]$.

3. Orthonormal System

Two quadratically integrable functions $\beta_i(y), \beta_j(y), y \in [a,b]$ are considered orthogonal to each other if the equality

$$\int_a^b \beta_i(y)\beta_j(y)\,dy = 0 \tag{11.44a}$$

holds. A system of functions $(\beta_n(y))$ in the space $L^2[a,b]$ is called an *orthonormal system* if the following equalities are true:

$$\int_a^b \beta_i(y)\beta_j(y)\,dy = \begin{cases} 1 & \text{for} \quad i=j, \\ 0 & \text{for} \quad i \neq j. \end{cases} \tag{11.44b}$$

An orthonormal system of functions is *complete* if there is no function $\tilde{\beta}(y) \neq 0$ in $L^2[a,b]$ orthogonal to every function of this system. A complete orthonormal system contains countably many functions. These functions form a *basis* of the space $L^2[a,b]$. To transform a system of functions $(\beta_n(y))$ into an orthonormal system $(\beta_n^*(y))$ the *Schmidt orthogonalization procedure* can be used. This determines the coefficients $b_{n1}, b_{n2}, \ldots, b_{nn}$ for $n = 1, 2, \ldots$ successively so that the function

$$\beta_n^*(y) = \sum_{j=1}^{n} b_{nj}\beta_j(y) \tag{11.44c}$$

is normalized and orthogonal to every function $\beta_1^*(y), \ldots, \beta_{n-1}^*(y)$.

4. Fourier Series

If $(\beta_n(y))$ is an orthonormal system and $\psi(y) \in L^2[a,b]$, then one calls the series

$$\sum_{j=1}^{\infty} d_j\beta_j(y) = \psi(y) \tag{11.45a}$$

the Fourier series of $\psi(y)$ with respect to $(\beta_n(y))$, and the numbers d_j are the corresponding Fourier coefficients. Based on (11.44b)

$$\int_a^b \beta_k(y)\psi(y)\,dy = \sum_{j=1}^{\infty} d_j \int_a^b \beta_j(y)\beta_k(y)\,dy = d_k \tag{11.45b}$$

holds. If $(\beta_n(y))$ is complete, the Parseval equality holds:

$$\int_a^b |\psi(y)|^2\,dy = \sum_{j=1}^{\infty} |d_j|^2. \tag{11.45c}$$

11.3.3 Reduction of an Integral Equation into a Linear System of Equations

A linear system of equations is needed in order to determine the Fourier coefficients of the solution function $\varphi(y)$ with respect to an orthonormal system. First, a complete orthonormal system $(\beta_n(y)), y \in [a,b]$ is chosen. A corresponding complete orthonormal system $(\alpha_n(x))$ can be chosen for the interval $x \in [c,d]$. With respect to the system $(\alpha_n(x))$ the function $f(x)$ has the Fourier series

$$f(x) = \sum_{i=1}^{\infty} f_i \alpha_i(x) \quad \text{with} \quad f_i = \int_c^d \alpha_i(x) f(x)\,dx. \tag{11.46a}$$

If the integral equation (11.41) is multiplied by $\alpha_i(x)$ and the integral is evaluated for x running from c to d alone, one gets:

$$\begin{aligned} f_i &= \int_c^d \int_a^b K(x,y)\varphi(y)\alpha_i(x)\,dy\,dx \\ &= \int_a^b \left\{ \int_c^d K(x,y)\alpha_i(x)\,dx \right\} \varphi(y)\,dy \quad (i = 1, 2, \ldots). \end{aligned} \tag{11.46b}$$

The expression in braces is a function of y with the Fourier representation

$$\int_c^d K(x,y)\alpha_i(x)\,dx = K_i(y) = \sum_{j=1}^{\infty} K_{ij}\beta_j(y) \quad \text{with} \tag{11.46c}$$

$$K_{ij} = \int_a^b \int_c^d K(x,y)\alpha_i(x)\beta_j(y)\,dx\,dy.$$

With the Fourier series approach

$$\varphi(y) = \sum_{k=1}^{\infty} c_k \beta_k(y) \tag{11.46d}$$

follows

$$\begin{aligned} f_i &= \int_a^b \left\{ \sum_{j=1}^{\infty} K_{ij}\beta_j(y) \left(\sum_{k=1}^{\infty} c_k \beta_k(y) \right) \right\} dy \\ &= \sum_{j=1}^{\infty} \sum_{k=1}^{\infty} K_{ij} c_k \int_a^b \beta_j(y)\beta_k(y)\,dy \quad (i = 1, 2, \ldots). \end{aligned} \tag{11.46e}$$

Because of the orthonormal property (11.44b) the system of equations

$$f_i = \sum_{j=1}^{\infty} K_{ij} c_j \quad (i = 1, 2, \ldots) \tag{11.46f}$$

holds. This is an infinite system of equations to determine the Fourier coefficients $c_1, c_2, \ldots$. The matrix of coefficients of the system of equations

$$\mathbf{K} = \begin{pmatrix} K_{11} & K_{12} & K_{13} & \cdots \\ K_{21} & K_{22} & K_{23} & \cdots \\ K_{31} & K_{32} & K_{33} & \cdots \\ \vdots & \vdots & \vdots & \end{pmatrix} \tag{11.46g}$$

is called a *kernel matrix*. The numbers f_i and K_{ij} $(i,j = 1,2,\dots)$ are known, although they depend on the orthonormal system chosen.

■ $f(x) = \dfrac{1}{\pi}\displaystyle\int_0^\pi \frac{\sin y}{\cos y - \cos x}\varphi(y)\,dy,\ \ 0 \le x \le \pi$. The integral is considered in the sense of the Cauchy principal value. As a complete orthogonal system can be used:

1. $\alpha_0(x) = \dfrac{1}{\sqrt{\pi}},\quad \alpha_i(x) = \sqrt{\dfrac{2}{\pi}}\cos ix\ \ (i = 1,2.\dots),$ **2.** $\beta_j(y) = \sqrt{\dfrac{2}{\pi}}\sin jy\ \ (j = 1,2,\dots).$

By (11.46d), the coefficients of the kernel matrix are

$$K_{0j} = \frac{1}{\sqrt{\pi}}\frac{1}{\pi}\sqrt{\frac{2}{\pi}}\int_0^\pi\int_0^\pi \frac{\sin y \sin jy}{\cos y - \cos x}\,dx\,dy = 0\ \ (j = 1,2,\dots),$$

$$K_{ij} = \frac{2}{\pi}\frac{1}{\pi}\int_0^\pi\int_0^\pi \frac{\sin y \sin iy \cos ix}{\cos y - \cos x}\,dx\,dy = \frac{2}{\pi^2}\int_0^\pi \sin y \sin iy\left\{\int_0^\pi \frac{\cos ix}{\cos y - \cos x}\,dx\right\}dy\ \ (i = 1,2,\dots).$$

For the inner integral the equation

$$\int_0^\pi \frac{\cos ix}{\cos y - \cos x}\,dx = -\pi\frac{\sin iy}{\sin y} \tag{11.47}$$

holds. Consequently $K_{ij} = -\dfrac{2}{\pi}\displaystyle\int_0^\pi \sin jy \sin iy\,dy = \begin{cases} 0 & \text{for } i \ne j, \\ -1 & \text{for } i = j. \end{cases}$

The Fourier coefficients of $f(x)$ from (11.46a) are $f_i = \displaystyle\int_0^\pi f(x)\alpha_i(x)\,dx\ \ (i = 0,1,2,\dots)$. The system of equations is $\begin{pmatrix} 0 & 0\ 0 & \cdots \\ -1 & 0\ 0 & \cdots \\ 0 & -1\ 0 & \cdots \\ \vdots & & \vdots \end{pmatrix}\begin{pmatrix} c_1 \\ c_2 \\ c_3 \\ \vdots \end{pmatrix} = \begin{pmatrix} f_0 \\ f_1 \\ f_2 \\ f_3 \\ \vdots \end{pmatrix}$. According to the first equation, the system can have any solution only if the equality $f_0 = \displaystyle\int_0^\pi f(x)\alpha_0(x)\,dx = \frac{1}{\sqrt{\pi}}\int_0^\pi f(x)\,dx = 0$ holds. Then $c_j = -f_j$ $(j = 1,2,\dots)$, and $\varphi(y) = -\sqrt{\dfrac{2}{\pi}}\displaystyle\sum_{j=1}^\infty f_j \sin jy = \frac{1}{\pi}\int_0^\pi \frac{\sin y}{\cos y - \cos x}f(x)\,dx$ holds.

11.3.4 Solution of the Homogeneous Integral Equation of the First Kind

If $\varphi(y)$ and $\varphi^h(y)$ are arbitrary solutions of the inhomogeneous and the homogeneous integral equation respectively, i.e.,

$$f(x) = \int_a^b K(x,y)\varphi(y)\,dy \tag{11.48a}$$

and

$$0 = \int_a^b K(x,y)\varphi^h(y)\,dy, \tag{11.48b}$$

then the sum $\varphi(y) + \varphi^h(y)$ is a solution of the inhomogeneous integral equation. Therefore there are to determine all the solutions of the homogeneous integral equation. This problem is the same as determining all the non-trivial solutions of the linear system of equations

$$\sum_{j=1}^\infty K_{ij}c_j = 0 \quad (i = 1,2,\dots). \tag{11.49}$$

As sometimes this system is not so easy to solve, the following method can be used for the calculations. If there is a complete orthonormal system $(\alpha_n(x))$, take the functions

$$K_i(y) = \int_c^d K(x,y)\alpha_i(x)\,dx \quad (i = 1, 2, \ldots). \tag{11.50a}$$

If $\varphi^h(y)$ is an arbitrary solution of the homogeneous equation, i.e.,

$$\int_a^b K(x,y)\varphi^h(y)\,dy = 0 \tag{11.50b}$$

holds, then multiplying this equality by $\alpha_i(x)$ and performing an integration with respect to x, gives

$$0 = \int_a^b \varphi^h(y) \int_c^d K(x,y)\alpha_i(x)\,dx\,dy = \int_a^b \varphi^h(y)K_i(y)\,dy \quad (i = 1, 2, \ldots), \tag{11.50c}$$

i.e., every solution $\varphi^h(y)$ of the homogeneous equation must be orthogonal to every function $K_i(y)$. Replacing the system $(K_n(y))$ by an orthonormal system $(K_n^*(y))$ and using an orthogonalization procedure, instead of (11.50c) follows

$$\int_a^b \varphi^h(y)K_i^*(y)\,dy = 0. \tag{11.50d}$$

Extending the system $(K_n^*(y))$ into a complete orthonormal system, the conditions (11.50d) are obviously valid for every linear combination of the new functions. If the orthonormal system $(K_n^*(y))$ is already complete, then only the trivial solution $\varphi^h(y) = 0$ exists.
The solution system of the adjoint homogeneous integral equation can be calculated in exactly the same way:

$$\int_c^d K(x,y)\psi(x)\,dx = 0. \tag{11.50e}$$

■ $\displaystyle\frac{1}{\pi}\int_0^\pi \frac{\sin x}{\cos y - \cos x}\varphi(y)\,dy = 0,\ 0 \le x \le \pi$. An orthonormal system is: $\alpha_i(x) = \sqrt{\frac{2}{\pi}}\sin ix$ $(i = 1, 2, \ldots)$, $K_i(y) = \sqrt{\frac{2}{\pi}}\frac{1}{\pi}\int_0^\pi \frac{\sin x \sin ix}{\cos y - \cos x}\,dx = \sqrt{\frac{2}{\pi}}\frac{1}{2\pi}\int_0^\pi \frac{\cos(i-1)x - \cos(i+1)x}{\cos y - \cos x}\,dx$. Applying (11.47) twice yields $K_i(y) = -\sqrt{\frac{2}{\pi}}\frac{1}{2}\left(\frac{\sin(i-1)y - \sin(i+1)y}{\sin y}\right) = \sqrt{\frac{2}{\pi}}\cos iy$ $(i = 1, 2, \ldots)$. The system $(K_n(y))$ is already an orthonormal system. The function $K_0(y) = \frac{1}{\sqrt{\pi}}$ completes this system.
Consequently the homogeneous equation has only the solution: $\varphi^h(y) = c\frac{1}{\sqrt{\pi}} = \tilde{c}$, ($c$ is arbitrary).

11.3.5 Construction of Two Special Orthonormal Systems for a Given Kernel

1. Preliminaries

The solution of infinite systems of linear equations (see 11.3.3, p. 638) is not usually easier than the solution of the original problem. Choosing suitable orthonormal systems $(\alpha_n(x))$ and $(\beta_n(y))$ one can change the structure of the kernel matrix $\mathbf{K}$ in such a way that the system of equations can be solved easily. By the following method two orthonormal systems can be constructed such that the coefficients

K_{ij} of the kernel matrix are non-zero only for $i = j$ and $i = j + 1$.

Using the method given in the previous paragraph, first two orthonormal systems $(\beta_n^h(y))$ and $(\alpha_n^h(x))$ are to determine, i.e., the solution systems of the homogeneous and the corresponding adjoint homogeneous integral equations respectively. This means that it can be given all the solutions of these two integral equations by a linear combination of the functions $\beta_n^h(y)$ and $\alpha_n^h(x)$. These orthonormal systems are not complete. By the following method these systems are to be completed step by step into complete orthonormal systems $\alpha_j(x), \beta_j(y) \ \ (j = 1, 2, \ldots)$.

2. Procedure

First a normalized function $\alpha_1(x)$ is determined, which is orthogonal to every function $\left(\alpha_n^h(x)\right)$. Then the following steps are performed for $j = 1, 2, \ldots$:
1. Determination of the function $\beta_j(y)$ and the number ν_j from the formulas

$$\nu_1\beta_1(y) = \int_c^d K(x,y)\alpha_1(x)\,dx \quad \text{or} \tag{11.51a}$$

$$\nu_j\beta_j(y) = \int_c^d K(x,y)\alpha_j(x)\,dx - \mu_{j-1}\beta_{j-1}(y) \quad (j \neq 1), \tag{11.51b}$$

so that ν_j is never equal to zero and $\beta_j(y)$ is normalized. Then $\beta_j(y)$ is orthogonal to the functions $\left((\beta_n^h(y)), \beta_1(y), \ldots, \beta_{j-1}(y)\right)$.

2. Determination of the function $\alpha_{j+1}(x)$ and the number μ_j from the formula

$$\mu_j\alpha_{j+1}(x) = \int_a^b K(x,y)\beta_j(y)\,dy - \nu_j\alpha_j(x). \tag{11.51c}$$

There are two possibilities:

a) $\mu_j \neq 0$: The function $\alpha_{j+1}(x)$ is orthogonal to the functions $\left((\alpha_n^h(x)), \alpha_1(x), \ldots, \alpha_j(x)\right)$.

b) $\mu_j = 0$: Then the function $\alpha_{j+1}(x)$ is not uniquely defined. Here again there are two cases:

$\mathbf{b_1}$) The system $\left((\alpha_n^h(x)), \alpha_1(x), \ldots, \alpha_j(x)\right)$ is already complete. Then the system $\left((\beta_n^h(y)), \beta_1(y), \ldots, \beta_j(y)\right)$ is also complete, and the procedure is finished.

$\mathbf{b_2}$) The system $\left((\alpha_n^h(x)), \alpha_1(x), \ldots, \alpha_j(x)\right)$ is not complete. Then again one chooses an arbitrary function $\alpha_{j+1}(x)$ orthogonal to the previous functions.

This procedure is repeated until the orthonormal systems are complete. It is possible that after a certain step the case **b)** does not occur during a countable number of steps, but the system of this countable number of functions $\left((\alpha_n^h(x)), \alpha_1(x), \ldots\right)$ is still not complete. Then again one can start the procedure by a function $\tilde{\alpha_1}(x)$, which is orthogonal to every function of the previous system.

If the functions $\alpha_j(x), \beta_j(y)$ and the numbers ν_j, μ_j are determined by the procedure given above, the kernel matrix $\mathbf{K}$ has the form

$$\mathbf{K} = \begin{pmatrix} 0 & 0 & 0 & \cdots \\ 0 & \mathbf{K^1} & 0 & \cdots \\ 0 & 0 & \mathbf{K^2} & \cdots \\ \vdots & & \cdots & \vdots \end{pmatrix} \quad \text{with} \quad \mathbf{K}^m = \begin{pmatrix} \nu_1^{(m)} & 0 & 0 & \cdots \\ \mu_1^{(m)} & \nu_2^{(m)} & 0 & \cdots \\ 0 & \mu_2^{(m)} & \nu_3^{(m)} & \cdots \\ \vdots & & \cdots & \vdots \end{pmatrix}. \tag{11.52}$$

The matrices $\mathbf{K}^m$ $(m = 1, 2, \ldots)$ are finite if during the procedure $\mu_j^{(m)} = 0$ holds after a finite number of steps. They are infinite if for countably many values of j, $\mu_j^{(m)} \neq 0$ holds. The number of zero rows and zero columns in $\mathbf{K}$ corresponds to the number of functions in the systems $(\alpha_n^h(x))$ and $(\beta_n^h(y))$. A very simple case happens if the matrices $\mathbf{K}^m$ contain one number $\nu_1^{(m)} = \nu_m$ only, i.e., all numbers $\mu_j^{(m)}$ are equal to zero.
Using the notation of 11.3.3, p. 638, for the solution of the infinite system of equations under the assumptions $f_j = 0$ for $\alpha_j(x) \in (\alpha_n^h(x))$ it holds:

$$c_j = \begin{cases} \dfrac{f_j}{\nu_j} & \text{for} \quad \beta_j(y) \notin \left(\beta_n^h(y)\right), \\ \text{arbitrary} & \text{for} \quad \beta_j(y) \in \left(\beta_n^h(y)\right). \end{cases} \tag{11.53}$$

11.3.6 Iteration Method

To solve the integral equation

$$f(x) = \int_a^b K(x,y)\varphi(y)\,dy \quad (c \le x \le d), \tag{11.54a}$$

starting with $\alpha_0(x) = f(x)$ one determines the functions

$$\beta_n(y) = \int_c^d K(x,y)\alpha_{n-1}(x)\,dx \quad (11.54\text{b}) \qquad \text{and} \quad \alpha_n(x) = \int_a^b K(x,y)\beta_n(y)\,dy, \tag{11.54c}$$

for $n = 1, 2, \ldots$. If there is a quadratically integrable solution $\varphi(y)$ of (11.54a) then the following equalities hold:

$$\begin{aligned} \int_a^b \varphi(y)\beta_n(y)\,dy &= \int_a^b \int_c^d \varphi(y)K(x,y)\alpha_{n-1}(x)\,dx\,dy \\ &= \int_c^d f(x)\alpha_{n-1}(x)\,dx \qquad (n = 1, 2, \ldots). \end{aligned} \tag{11.54d}$$

Orthogonalization and normalization of the function systems (11.54b,c) give the orthonormal systems $(\alpha_n^*(x))$ and $(\beta_n^*(y))$. Using the Schmidt orthogonalization method then $\beta_n^*(y)$ has the form

$$\beta_n^*(y) = \sum_{j=1}^{n} b_{nj}\beta_j(y) \qquad (n = 1, 2, \ldots). \tag{11.54e}$$

Now it is assumed that the solution $\varphi(y)$ of (11.54a) has the representation by the series

$$\varphi(y) = \sum_{j=1}^{\infty} c_n\beta_n^*(y). \tag{11.54f}$$

In this case for the coefficients c_n regarding (11.54d)

$$c_n = \int_a^b \varphi(y)\beta_n^*(y)\,dy = \sum_{j=1}^{n} b_{nj} \int_a^b \varphi(y)\beta_j(y)\,dy = \sum_{j=1}^{n} b_{nj} \int_c^d f(x)\alpha_{j-1}(x)\,dx. \tag{11.54g}$$

holds. To have a solution in the form (11.54f) the following conditions are both necessary and sufficient:

1. $\int_c^d [f(x)]^2\,dx = \sum_{n=1}^{\infty} \Big| \int_c^d f(x)\alpha_n^*(x)\,dx \Big|^2$, (11.55a) **2.** $\sum_{n=1}^{\infty} |c_n|^2 < \infty.$ (11.55b)

11.4 Volterra Integral Equations

11.4.1 Theoretical Foundations

A Volterra integral equation of the second kind has the form

$$\varphi(x) = f(x) + \int_a^x K(x,y)\varphi(y)\,dy. \tag{11.56}$$

The solution function $\varphi(x)$ with the independent variable x from the closed interval $I = [a,b]$ or from the semi-open interval $I = [a,\infty)$ is required. There is the following theorem about the solution of a Volterra integral equation of the second kind: If the functions $f(x)$ for $x \in I$ and $K(x,y)$ on the triangular region $x \in I$ and $y \in [a,x]$ are continuous, then there exists a *unique* solution $\varphi(x)$ of the integral equation such that it is continuous for $x \in I$. For this solution

$$\varphi(a) = f(a) \tag{11.57}$$

holds. In many cases, the Volterra integral equation of the first kind can be transformed into an equation of the second kind. Hence, theorems about existence and uniqueness of the solution are valid with some modifications.

1. Transformation by Differentiation

Assuming $\varphi(x)$, $K(x,y)$, and $K_x(x,y)$ are continuous functions, the integral equation of the first kind

$$f(x) = \int_a^x K(x,y)\varphi(y)\,dy \tag{11.58a}$$

can be transformed into the form

$$f'(x) = K(x,x)\varphi(x) + \int_a^x \frac{\partial}{\partial x}K(x,y)\varphi(y)\,dy \tag{11.58b}$$

by differentiation with respect to x. If $K(x,x) \neq 0$ for all $x \in I$, then dividing the equation by $K(x,x)$ gives an integral equation of the second kind.

2. Transformation by Partial Integration

Assuming that $\varphi(x)$, $K(x,y)$ and $K_y(x,y)$ are continuous, one can evaluate the integral in (11.58a) by partial integration. Substituting

$$\int_a^x \varphi(y)\,dy = \psi(x) \tag{11.59a}$$

gives

$$\begin{aligned} f(x) &= [K(x,y)\psi(y)]_{y=a}^{y=x} - \int_a^x \left(\frac{\partial}{\partial y}K(x,y)\right)\psi(y)\,dy \\ &= K(x,x)\psi(x) - \int_a^x \left(\frac{\partial}{\partial y}K(x,y)\right)\psi(y)\,dy. \end{aligned} \tag{11.59b}$$

If $K(x,x) \neq 0$ for $x \in I$, then dividing by $K(x,x)$ gives an integral equation of the second kind:

$$\psi(x) = \frac{f(x)}{K(x,x)} + \frac{1}{K(x,x)} \int_a^x \left(\frac{\partial}{\partial y} K(x,y) \right) \psi(y)\, dy. \tag{11.59c}$$

Differentiating the solution $\psi(x)$ yields the solution $\varphi(x)$ of (11.58a).

11.4.2 Solution by Differentiation

In some Volterra integral equations the integral vanishes after differentiation with respect to x, or it can be suitably substituted. Assuming that the functions $K(x,y)$, $K_x(x,y)$, and $\varphi(x)$ are continuous or, in the case of an integral equation of the second kind, $\varphi(x)$ is differentiable, and differentiating

$$f(x) = \int_a^x K(x,y)\varphi(y)\, dy \quad (11.60a) \qquad \text{or} \quad \varphi(x) = f(x) + \int_a^x K(x,y)\varphi(y)\, dy \tag{11.60b}$$

with respect to x yields

$$f'(x) = K(x,x)\varphi(x) + \int_a^x \frac{\partial}{\partial x} K(x,y)\, \varphi(y)\, dy \quad \text{or} \tag{11.60c}$$

$$\varphi'(x) = f'(x) + K(x,x)\varphi(x) + \int_a^x \frac{\partial}{\partial x} K(x,y)\, \varphi(y)\, dy. \tag{11.60d}$$

■ Find the solution $\varphi(x)$ for $x \in \left[0, \frac{\pi}{2}\right)$ of the equation $\int_0^x \cos(x-2y)\varphi(y)\, dy = \frac{1}{2} x \sin x$ (I). Differentiating it twice with respect to x gives $\varphi(x)\cos x - \int_0^x \sin(x-2y)\varphi(y)\, dy = \frac{1}{2}(\sin x + x\cos x)$ (II a), and $\varphi'(x)\cos x - \int_0^x \cos(x-2y)\varphi(y)\, dy = \cos x - \frac{1}{2} x \sin x$ (IIb). The integral in the second equation is the same as that in the original problem, so one can substitute it. This yields $\varphi'(x)\cos x = \cos x$ and because $\cos x \neq 0$ for $x \in \left[0, \frac{\pi}{2}\right)$, $\varphi'(x) = 1, \quad \text{so} \quad \varphi(x) = x + C$.

To determine the constant C substitute $x = 0$ in (IIa) to obtain $\varphi(0) = 0$. Consequently $C = 0$, and the solution of (I) is $\varphi(x) = x$.

Remark: If the kernel of a Volterra integral equation is a polynomial, then one can transform the integral equation by differentiation into a linear differential equation. Suppose the highest power of x in the kernel is n. After differentiating the equation $(n+1)$ times with respect to x follows a differential equation of n-th order in the case of an integral equation of the first kind, and of the order $n+1$ in the case of an integral equation of the second kind. Of course it is to be assumed that $\varphi(x)$ and $f(x)$ are differentiable as many times as necessary.

■ $\int_0^x [2(x-y)^2 + 1]\varphi(y)\, dy = x^3$ (I*). After differentiating three times with respect to x holds $\varphi(x) + 4\int_0^x (x-y)\varphi(y)\, dy = 3x^2$ (II*a), $\varphi'(x) + 4\int_0^x \varphi(y)\, dy = 6x$ (II*b), $\varphi''(x) + 4\varphi(x) = 6$ (II*c). The general solution of this differential equation is $\varphi(x) = A\sin 2x + B\cos 2x + \frac{3}{2}$. Substituting $x = 0$ in (II*a) and (II*b) results in $\varphi(0) = 0$, $\varphi'(0) = 0$, so the solution is $A = 0$, $B = -1.5$. The solution of the integral equation (I*) is $\varphi(x) = \frac{3}{2}(1 - \cos 2x)$.

11.4.3 Solution of the Volterra Integral Equation of the Second Kind by Neumann Series

The solution of the Volterra integral equations of the second kind can be represented by using a Neumann series (see 11.2.3, p. 627). If the equation has the form

$$\varphi(x) = f(x) + \lambda \int_a^x K(x,y)\varphi(y)\,dy, \tag{11.61}$$

so one formally substitutes

$$\overline{K}(x,y) = \begin{cases} K(x,y) & \text{for } y \le x, \\ 0 & \text{for } y > x. \end{cases} \tag{11.62a}$$

With this transformation (11.61) is identically to a Fredholm integral equation

$$\varphi(x) = f(x) + \lambda \int_a^b \overline{K}(x,y)\varphi(y)\,dy, \tag{11.62b}$$

allowing $b = \infty$ as well. The solution has the representation

$$\varphi(x) = f(x) + \sum_{n=1}^{\infty} \lambda^n \int_a^b K_n(x,y) f(y)\,dy. \tag{11.62c}$$

The *iterated kernels* $K_1, K_2, \ldots$ are defined by the following equalities:

$$K_1(x,y) = \overline{K}(x,y), \quad K_2(x,y) = \int_a^b \overline{K}(x,\eta)\overline{K}(\eta,y)\,d\eta = \int_y^x K(x,\eta)K(\eta,y)\,d\eta, \ldots \tag{11.62d}$$

and in general:

$$K_n(x,y) = \int_y^x K(x,\eta)K_{n-1}(\eta,y)\,d\eta. \tag{11.62e}$$

The equalities $K_j(x,y) \equiv 0$ for $y > x$ $(j = 1,2,\ldots)$ are also valid for iterated kernels. Contrary to Fredholm integral equations if (11.61) has any solution, the Neumann series converges to it regardless of the value of λ.

■ $\varphi(x) = 1 + \lambda \int_0^x e^{x-y}\varphi(y)\,dy$. $K_1(x,y) = K(x,y) = e^{x-y}$, $K_2(x,y) = \int_y^x e^{x-\eta}e^{\eta-y}\,d\eta = e^{x-y}(x-y), \ldots,$ $K_n(x,y) = \dfrac{e^{x-y}}{(n-1)!}(x-y)^{n-1}$.

Consequently the resolvent is: $\Gamma(x,y;\lambda) = e^{x-y}\sum_{n=0}^{\infty}\dfrac{\lambda^n}{n!}(x-y)^n = e^{(x-y)(\lambda+1)}$. It is well-known that this series is convergent for any value of the parameter λ.

One gets $\varphi(x) = 1 + \lambda\int_0^x e^{(x-y)(\lambda+1)}\,dy = 1 + \lambda e^{(\lambda+1)x}\int_0^x e^{-(\lambda+1)y}\,dy$, in particular if $\lambda = -1$: $\varphi(x) = 1 - x$, $\lambda \neq -1$: $\varphi(x) = \dfrac{1}{\lambda+1}\left(1 + \lambda e^{(\lambda+1)x}\right)$.

11.4.4 Convolution Type Volterra Integral Equations

If the kernel of a Volterra integral equation has the special form

$$K(x,y) = \begin{cases} k(x-y) & \text{for } 0 \le y \le x, \\ 0 & \text{for } 0 \le x < y, \end{cases} \tag{11.63a}$$

can be used the Laplace transformation to solve the equations

$$\int_0^x k(x-y)\varphi(y)\,dy = f(x) \quad (11.63b) \qquad \text{or} \quad \varphi(x) = f(x) + \int_0^x k(x-y)\varphi(y)\,dy. \quad (11.63c)$$

If the Laplace transforms $\mathcal{L}\{\varphi(x)\} = \Phi(p)$, $\mathcal{L}\{f(x)\} = F(p)$, and $\mathcal{L}\{k(x)\} = K(p)$ exist, then the transformed equations have the form (see 15.2.1.2, **11.**, p. 773)

$$K(p)\Phi(p) = F(p) \quad (11.64a) \quad \text{or} \quad \Phi(p) = F(p) + K(p)\Phi(p) \ \text{resp.} \quad (11.64b)$$

From these follows

$$\Phi(p) = \frac{F(p)}{K(p)} \quad (11.64c) \quad \text{or} \quad \Phi(p) = \frac{F(p)}{1-K(p)} \ \text{resp.} \quad (11.64d)$$

The inverse transformation gives the solution $\varphi(x)$ of the original problem. Rewriting the formula for the Laplace transform of the solution of the integral equation of the second kind yields

$$\Phi(p) = \frac{F(p)}{1-K(p)} = F(p) + \frac{K(p)}{1-K(p)}F(p). \quad (11.64e)$$

The formula

$$\frac{K(p)}{1-K(p)} = H(p) \quad (11.64f)$$

depends only on the kernel, and denoting its inverse by $h(x)$, the solution is

$$\varphi(x) = f(x) + \int_0^x h(x-y)f(y)\,dy. \quad (11.64g)$$

The function $h(x-y)$ is the resolvent kernel of the integral equation.

■ $\varphi(x) = f(x) + \int_0^x e^{x-y}\varphi(y)\,dy$: $\Phi(p) = F(p) + \frac{1}{p-1}\Phi(p)$, i.e., $\Phi(p) = \frac{p-1}{p-2}F(p)$. The inverse transformation gives $\varphi(x)$. From $H(p) = \frac{1}{p-2}$ it follows that $h(x) = e^{2x}$. By (11.64g) the solution is $\varphi(x) = f(x) + \int_0^x e^{2(x-y)}f(y)\,dy$.

11.4.5 Numerical Methods for Volterra Integral Equation of the Second Kind

The problem is to find the solution for the integral equation

$$\varphi(x) = f(x) + \int_a^x K(x,y)\varphi(y)\,dy \quad (11.65)$$

for x from the interval $I = [a,b]$. The purpose of numerical methods is somehow to approximate the integral by a quadrature formula:

$$\int_a^x K(x,y)\varphi(y)\,dy \approx Q_{[a,x]}(K(x,.)\varphi(.)). \quad (11.66a)$$

Both the interval of integration and the quadrature formula depend on x. This fact is emphasized by the index $[a,x]$ of $Q_{[a,x]}(\ldots)$. One gets the following equation as an approximation of (11.65):

$$\overline{\varphi}(x) = f(x) + Q_{[a,x]}(K(x,.)\overline{\varphi}(.)). \quad (11.66b)$$

The function $\overline{\varphi}(x)$ is an approximation of the solution of (11.65). The number and the arrangement of the interpolation nodes of the quadrature formula depend on x, so as to allow little choice. If ξ is an interpolation node of $Q_{[a,x]}(K(x,.)\overline{\varphi}(.))$, then $(K(x,\xi)\overline{\varphi}(\xi))$ and especially $\overline{\varphi}(\xi)$ must be known. For this purpose, the right-hand side of (11.66b) should be evaluated first for $x = \xi$, which is equivalent to a quadrature over the interval $[a,\xi]$. As a consequence, the use of the popular Gauss quadrature formula is not possible.

The problem is to be solved by choosing the interpolation nodes as $a = x_0 < x_1 < \ldots < x_k < \ldots$ and using a quadrature formula $Q_{[a,x_n]}$ with the interpolation nodes $x_0, x_1, \ldots, x_n$. The values of the function at the interpolation nodes are denoted by the brief notation $\varphi_k = \overline{\varphi}(x_k) \quad (k = 0,1,2,\ldots)$. For φ_0 follows (see 11.3.1, p. 635)

$$\varphi_0 = f(x_0) = f(a), \qquad (11.66c) \qquad \text{and with this:} \quad \varphi_1 = f(x_1) + Q_{[a,x_1]}(K(x_1,.)\overline{\varphi}(.)). \qquad (11.66d)$$

$Q_{[a,x_1]}$ has the interpolation points x_0 and x_1 and consequently it has the form

$$Q_{[a,x_1]}(K(x_1,.)\overline{\varphi}(.) = w_0 K(x_1,x_0)\varphi_0 + w_1 K(x_1,x_1)\varphi_1 \qquad (11.66e)$$

with suitable coefficients w_0 and w_1. Continuing this procedure, the values φ_k are successively determined from the general relation:

$$\varphi_k = f(x_k) + Q_{[a,x_k]}(K(x_k,.)\overline{\varphi}(.)), \quad k = 1,2,3,\ldots . \qquad (11.66f)$$

The quadrature formulas $Q_{[a,x_k]}$ have the following form:

$$Q_{[a,x_k]}(K(x_k,.)\overline{\varphi}(.) = \sum_{j=0}^{k} w_{jk} K(x_k,x_j)\varphi_j. \qquad (11.66g)$$

Hence, (11.66f) takes the form:

$$\varphi_k = f(x_k) + \sum_{j=0}^{k} w_{jk} K(x_k,x_j)\varphi_j. \qquad (11.66h)$$

The simplest quadrature formula is the *left-hand rectangular formula* (see 19.3.2.1, p. 964). For this the coefficients are

$$w_{jk} = x_{j+1} - x_j \quad \text{for } j < k \quad \text{and} \quad w_{kk} = 0. \qquad (11.66i)$$

With this follows the system

$$\begin{aligned}
\varphi_0 &= f(a), \\
\varphi_1 &= f(x_1) + (x_1 - x_0)K(x_1,x_0)\varphi_0, \qquad (11.67a)\\
\varphi_2 &= f(x_2) + (x_1 - x_0)K(x_2,x_0)\varphi_0 + (x_2 - x_1)K(x_2,x_1)\varphi_1
\end{aligned}$$

and generally

$$\varphi_k = f(x_k) + \sum_{j=0}^{k-1} (x_{j+1} - x_j)K(x_k,x_j)\varphi_j. \qquad (11.67b)$$

More accurate approximations of the integral can be obtained by using the *trapezoidal formula* (see 19.3.2.2, p. 964). To make it simple, one chooses equidistant interpolation nodes $x_k = a + kh, \;\; k = 0,1,2,\ldots$:

$$\int_a^b g(x)\,dx \approx \frac{h}{2}\left[g(x_0) + 2\sum_{j=1}^{k-1} g(x_j) + g(x_k)\right]. \qquad (11.67c)$$

Using this approximation for (11.66f) one gets:

$$\varphi_0 = f(a), \qquad (11.67d)$$

$$\varphi_k = f(x_k) + \frac{h}{2}\left[K(x_k,x_0)\varphi_0 + K(x_k,x_k)\varphi_k + 2\sum_{j=1}^{k-1} K(x_k,x_j)\varphi_j\right]. \qquad (11.67e)$$

Although the unknown values also appear on the right-hand side of the equation, they are easy to express.

Remark: With the previous method one can approximate the solution of non-linear integral equations as well. Using the trapezoidal formula to determine the values φ_k one has to solve a non-linear equation. To avoid this one can use the trapezoidal formula for the interval $[a, x_{k-1}]$, and use the rectangular formula for the interval $[x_{k-1}, x_k]$. If h is small enough, this quadrature error does not have a significant effect on the solution.

■ The problem is to solve the integral equation $\varphi(x) = 2 + \int_0^x (x-y)\varphi(y)\,dy$ by the formula (11.66f) using the rectangular formula. The interpolation nodes are the equidistant points $x_k = k \cdot 0.1$, and hence $h = 0.1$.

$$\begin{aligned}
\varphi_0 &= 2, \\
\varphi_1 &= f(x_1) + hK(x_1, x_0)\,\varphi_0 \\
&= 2 + 0.1 \cdot 0.1 \cdot 2 = 2.02, \\
\varphi_2 &= f(x_2) + h(K(x_2, x_0)\,\varphi_0 + K(x_2, x_1)\,\varphi_1) \\
&= 2 + 0.1(0.2 \cdot 2 + 0.1 \cdot 2.02) = 2.0602
\end{aligned}$$

etc.

x	exact	rectangular formula	trapezoidal formula
0.2	2.0401	2.0602	2.0401
0.4	2.1621	2.2030	2.1620
0.6	2.3709	2.4342	2.3706
0.8	2.6749	2.7629	2.6743
1.0	3.0862	3.2025	3.0852

In the table the values of the exact solution are given, as well as the approximate values calculated by the rectangular and the trapezoidal formulas, respectively, so the accuracies of these methods can be compared. The step size used is $h = 0.1$.

11.5 Singular Integral Equations

An integral equation is called a *singular integral equation* if the range of the integral in the equation is not finite, or if the kernel has singularities inside of the range of integration. It is to be supposed that the integrals exist as improper integrals, or as Cauchy principal values (see 8.2.3, p. 506ff.). The properties and the conditions for the solutions of singular integral equations are very different from those in the case of "ordinary" integral equations. In the following sections only some special problems are discussed. For further discussions see [11.2], [11.3], [11.7], [11.8].

11.5.1 Abel Integral Equation

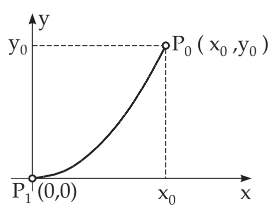


Figure 11.2

One of the first applications of integral equations for a physical problem was considered by Abel. A particle is moving in a vertical plane along a curve under the influence only of gravity from the point $P_0(x_0, y_0)$ to the point $P_1(0, 0)$ (**Fig. 11.2**).
The velocity of the particle at a point of the curve is

$$v = \frac{ds}{dt} = \sqrt{2g(y_0 - y)}\,. \tag{11.68}$$

The time of fall as a function of y_0 is calculated by integration:

$$T(y_0) = \int_0^l \frac{ds}{\sqrt{2g(y_0 - y)}}\,. \tag{11.69a}$$

If s is considered as a function of y, i.e., $s = f(y)$, then

$$T(y_0) = \int_0^{y_0} \frac{1}{\sqrt{2g}} \cdot \frac{f'(y)}{\sqrt{y_0 - y}}\,dy. \tag{11.69b}$$

The next problem is to determine the shape of the curve as a function of y_0 if the time of the fall is given. By substitution

$$\sqrt{2g}\cdot T(y_0)=F(y_0) \quad \text{and} \quad f'(y)=\varphi(y) \tag{11.69c}$$

and changing the notation of the variable y_0 into x, a Volterra integral equation of the first kind is obtained:

$$F(x)=\int_0^x \frac{\varphi(y)}{\sqrt{x-y}}\,dy. \tag{11.69d}$$

It is to be considered now the slightly more general equation

$$f(x)=\int_a^x \frac{\varphi(y)}{(x-y)^\alpha}\,dy \quad \text{with} \quad 0<\alpha<1. \tag{11.70}$$

The kernel of this equation is not bounded for $y=x$. In (11.70), the variable y is formally replaced by ξ and the variable x by y. By these substitutions the solution is obtained in the form $\varphi=\varphi(x)$. If both sides of (11.70) are multiplied by the term $\dfrac{1}{(x-y)^{1-\alpha}}$ and integrated with respect to y between the limits a and x, it yields the equation

$$\int_a^x \frac{1}{(x-y)^{1-\alpha}}\left(\int_a^y \frac{\varphi(\xi)}{(y-\xi)^\alpha}\,d\xi\right)dy=\int_a^x \frac{f(y)}{(x-y)^{1-\alpha}}\,dy. \tag{11.71a}$$

Changing the order of integration on the left-hand side gives

$$\int_a^x \varphi(\xi)\left\{\int_\xi^x \frac{dy}{(x-y)^{1-\alpha}(y-\xi)^\alpha}\right\}d\xi=\int_a^x \frac{f(y)}{(x-y)^{1-\alpha}}\,dy. \tag{11.71b}$$

The inner integral can be evaluated by the substitution $y=\xi+(x-\xi)u$:

$$\int_\xi^x \frac{dy}{(x-y)^{1-\alpha}(y-\xi)^\alpha}=\int_0^1 \frac{du}{u^\alpha(1-u)^{1-\alpha}}=\frac{\pi}{\sin(\alpha\pi)}\,. \tag{11.71c}$$

This result is to be substituted into (11.71b). After differentiation with respect to x one gets the function $\varphi(x)$:

$$\varphi(x)=\frac{\sin(\alpha\pi)}{\pi}\frac{d}{dx}\int_a^x \frac{f(y)}{(x-y)^{1-\alpha}}\,dy. \tag{11.71d}$$

■ $x=\int_0^x \dfrac{\varphi(y)}{\sqrt{x-y}}\,dy, \quad \varphi(x)=\dfrac{1}{\pi}\dfrac{d}{dx}\int_0^x \dfrac{y}{\sqrt{x-y}}\,dy=\dfrac{2}{\pi}\sqrt{x}\,.$

11.5.2 Singular Integral Equation with Cauchy Kernel

11.5.2.1 Formulation of the Problem

Consider the following integral equation:

$$a(x)\varphi(x)+\frac{1}{\pi\,\mathrm{i}}\int_\Gamma \frac{K(x,y)}{y-x}\varphi(y)\,dy=f(x), \quad x\in\Gamma. \tag{11.72}$$

Γ is a system consisting of a finite number of smooth, simple closed curves in the complex plane such that they form a connected interior domain S^+ with $0\in S^+$ and an exterior domain S^-. Driving along

the curve there is S^+ always on the left-hand side of Γ. A function $u(x)$ is *Hölder continuous* (or satisfies the *Hölder condition*) over Γ if for any pair $x_1,\ x_2 \in \Gamma$ the relations

$$|u(x_1) - u(x_2)| < K|x_1 - x_2|^\beta, \quad 0 < \beta \le 1, \quad K > 0 \tag{11.73}$$

are valid. It is supposed that the functions $a(x)$, $f(x)$, and $\varphi(x)$ are Hölder continuous with exponent β_1, and $K(x,y)$ is Hölder continuous with respect to both variables with the exponents $\beta_2 > \beta_1$. The kernel $K(x,y)(y-x)^{-1}$ has a strong singularity for $x = y$. The integral exists as a Cauchy principal value. With $K(x,x) = b(x)$ and $k(x,y) = \dfrac{K(x,y) - K(x,x)}{y - x}$ (11.72) holds in the form

$$(\mathcal{L}\varphi)(x) := a(x)\varphi(x) + \frac{b(x)}{\pi\,\mathrm{i}}\int_\Gamma \frac{\varphi(y)}{y-x}\,dy + \frac{1}{\pi\,\mathrm{i}}\int_\Gamma k(x,y)\varphi(y)\,dy = f(x), \quad x \in \Gamma. \tag{11.74a}$$

The expression $(\mathcal{L}\varphi)(x)$ denotes the left-hand side of the integral equation in abbreviated form. $\mathcal{L}$ is a singular operator. The function $k(x,y)$ is a weakly singular kernel. It is assumed that the normality condition $a(x)^2 - b(x)^2 \ne 0$, $x \in \Gamma$ holds. The equation

$$(\mathcal{L}_0\varphi)(x) = a(x)\varphi(x) + \frac{b(x)}{\pi\,\mathrm{i}}\int_\Gamma \frac{\varphi(y)}{y-x}dy = f(x), \quad x \in \Gamma, \tag{11.74b}$$

is the *characteristic equation* pertaining to (11.74a). The operator $\mathcal{L}_0$ is the characteristic part of the operator $\mathcal{L}$. The adjoint integral equation of (11.74a) yields the equality

$$\begin{aligned}(\mathcal{L}^\top\psi)(y) &= a(y)\psi(y) - \frac{b(y)}{\pi\,\mathrm{i}}\int_\Gamma \frac{\psi(x)dx}{x-y} + \frac{1}{\pi\,\mathrm{i}}\int_\Gamma \left(k(x,y) - \frac{b(x)-b(y)}{x-y}\right)\psi(x)\,dx \\ &= g(y), \quad y \in \Gamma.\end{aligned} \tag{11.74c}$$

11.5.2.2 Existence of a Solution

The equation $(\mathcal{L}\varphi)(x) = f(x)$ has a solution $\varphi(x)$ if and only if for every solution $\psi(y)$ of the homogeneous transposed equation $(\mathcal{L}^\top\psi)(y) = 0$ the condition of orthogonality

$$\int_\Gamma f(y)\psi(y)\,dy = 0 \tag{11.75a}$$

is satisfied. Similarly, the transposed equation $(\mathcal{L}^\top\psi)(y) = g(y)$ has a solution if for every solution $\varphi(x)$ of the homogeneous equation $(\mathcal{L}\varphi)(x) = 0$ the following is valid:

$$\int_\Gamma g(x)\varphi(x)\,dx = 0. \tag{11.75b}$$

11.5.2.3 Properties of Cauchy Type Integrals

The following function is called *a Cauchy type integral* over Γ:

$$\Phi(z) = \frac{1}{2\pi\,\mathrm{i}}\int_\Gamma \frac{\varphi(y)}{y-z}\,dy, \quad z \in \mathbb{C}, \tag{11.76a}$$

For $z \notin \Gamma$ the integral exists in the usual sense and represents a holomorphic function (see 14.1.2, p. 732). Also $\Phi(\infty) = 0$ holds. For $z = x \in \Gamma$ in (11.76a) is to be considered the Cauchy principal value

$$(\mathcal{H}\varphi)(x) = \frac{1}{2\pi\,\mathrm{i}}\int_\Gamma \frac{\varphi(y)}{y-x}\,dy, \quad x \in \Gamma. \tag{11.76b}$$

The Cauchy type integral $\Phi(z)$ can be extended continuously over Γ from S^+ and from S^-. The limits when approaching z the point $x \in \Gamma$ are denoted by $\Phi^+(x)$ and $\Phi^-(x)$, respectively. The formulas of Plemelj and Sochozki are valid:

$$\Phi^+(x) = \frac{1}{2}\varphi(x) + (\mathcal{H}\varphi)(x), \qquad \Phi^-(x) = -\frac{1}{2}\varphi(x) + (\mathcal{H}\varphi)(x). \tag{11.76c}$$

11.5.2.4 The Hilbert Boundary Value Problem

1. Relations

The solution of the characteristic integral equation and the Hilbert boundary value problem are strongly correlate. If $\varphi(x)$ is a solution of (11.74b), then (11.76a) is a holomorphic function on S^+ and S^- with $\Phi(\infty) = 0$. Because of the formulas of Plemelj and Sochozki (11.76c)

$$\varphi(x) = \Phi^+(x) - \Phi^-(x), \quad 2(\mathcal{H}\varphi)(x) = \Phi^+(x) + \Phi^-(x), \quad x \in \Gamma \tag{11.77a}$$

holds. With the notation

$$G(x) = \frac{a(x) - b(x)}{a(x) + b(x)} \quad \text{and} \quad g(x) = \frac{f(x)}{a(x) + b(x)}\,, \tag{11.77b}$$

the characteristic integral equation has the form:

$$\Phi^+(x) = G(x)\Phi^-(x) + g(x), \quad x \in \Gamma. \tag{11.77c}$$

2. Hilbert Boundary Value Problem

Looking for a function $\Phi(z)$ which is holomorphic on S^+ and S^-, and vanishes at infinity, and satisfies the boundary conditions (11.77c) over Γ. A solution $\Phi(z)$ of the Hilbert problem can be given in the form (11.76a). So, as a consequence of the first equation of (11.77a), a solution $\varphi(x)$ of the characteristic integral equation is determined.

11.5.2.5 Solution of the Hilbert Boundary Value Problem (in short: Hilbert Problem)

1. Homogeneous Boundary Conditions

$$\Phi^+(x) = G(x)\Phi^-(x), \quad x \in \Gamma. \tag{11.78}$$

During a single circulation of the point x along the curve Γ_l the value of $\log G(x)$ changes by $2\pi\,\mathrm{i}\lambda_l$, where λ_l is an integer. The change of the value of the function $\log G(x)$ during a single traverse of the complete curve system Γ is

$$\sum_{l=0}^{n} 2\pi\,\mathrm{i}\,\lambda_l = 2\pi\,\mathrm{i}\,\kappa. \tag{11.79a}$$

The number $\kappa = \sum\limits_{l=0}^{n} \lambda_l$ is called the *index of the Hilbert problem.* Now it is to compose a function

$$G_0(x) = (x - a_0)^{-\kappa}\Pi(x)G(x) \tag{11.79b}$$

$$\Pi(x) = (x - a_1)^{\lambda_1}(x - a_2)^{\lambda_2} \cdots (x - a_n)^{\lambda_n}, \tag{11.79c}$$

where $a_0 \in S^+$ and a_l $(l = 1, \dots, n)$ are arbitrarily fixed points inside Γ_l. If $\Gamma = \Gamma_0$ is a simple closed curve $(n = 0)$, then one defines $\Pi(x) = 1$. With

$$I(z) := \frac{1}{2\pi\,\mathrm{i}} \int\limits_{\Gamma} \frac{\log G_0(y)}{y - z}\, dy \tag{11.79d}$$

the following particular solution of the homogeneous Hilbert problems is obtained, which is called the fundamental solution:

$$X(z) = \begin{cases} \Pi^{-1}(z)\exp I(z) & \text{for } z \in S^+, \\ (z - a_0)^{-\kappa}\exp I(z) & \text{for } z \in S^-. \end{cases} \tag{11.79e}$$

This function doesn't vanish for any finite z. The most general solution of the homogeneous Hilbert problem, which vanishes at infinity, for $\kappa > 0$ is

$$\Phi_h(z) = X(z)P_{\kappa-1}(z), \quad z \in \mathbb{C} \tag{11.80}$$

with an arbitrary polynomial $P_{\kappa-1}(z)$ of degree at most $(\kappa - 1)$. For $\kappa \le 0$ there exists only the trivial solution $\Phi_h(z) = 0$ which satisfies the condition $\Phi_h(\infty) = 0$, so in this case $P_{\kappa-1}(z) \equiv 0$. For $\kappa > 0$ the homogeneous Hilbert problem has κ linearly independent solutions vanishing at infinity.

2. Inhomogeneous Boundary Conditions

The solution of the inhomogeneous Hilbert problem is the following:

$$\Phi(z) = X(z)R(z) + \Phi_h(z) \tag{11.81}$$ with $$R(z) = \frac{1}{2\pi\,\mathrm{i}} \int\limits_{\Gamma} \frac{g(y)dy}{X^+(y)(y-z)}\,. \tag{11.82}$$

If $\kappa < 0$ holds, for the existence of a solution vanishing at infinity the following necessary and sufficient conditions must be fulfilled:

$$\int\limits_{\Gamma} \frac{y^k g(y)dy}{X^+(y)} = 0 \quad (k = 0, 1, \ldots, -\kappa - 1). \tag{11.83}$$

11.5.2.6 Solution of the Characteristic Integral Equation

1. Homogeneous Characteristic Integral Equation

If $\Phi_h(z)$ is the solution of the corresponding homogeneous Hilbert problem, from (11.77a) follows the solution of the homogeneous integral equation

$$\varphi_h(x) = \Phi_h^+(x) - \Phi_h^-(x), \quad x \in \Gamma. \tag{11.84a}$$

For $\kappa \le 0$ only the trivial solution $\varphi_h(x) = 0$ exists. For $\kappa > 0$ the general solution is

$$\varphi_h(x) = [X^+(x) - X^-(x)]P_{\kappa-1}(x) \tag{11.84b}$$

with a polynomial $P_{\kappa-1}$ of degree at most $\kappa - 1$.

2. Inhomogeneous Characteristic Integral Equation

If $\Phi(z)$ is a general solution of the inhomogeneous Hilbert problem, the solution of the inhomogeneous integral equation can be given by (11.77a):

$$\varphi(x) = \Phi^+(x) - \Phi^-(x) \tag{11.85a}$$

$$= X^+(x)R^+(x) - X^-(x)R^-(x) + \Phi_h^+(x) - \Phi_h^-(x), \quad x \in \Gamma. \tag{11.85b}$$

Using the formulas of Plemelj and Sochozki (11.76c) for $R(z)$ holds

$$R^+(x) = \frac{1}{2}\frac{g(x)}{X^+(x)} + \left(\mathcal{H}\frac{g}{X^+}\right)(x), \quad R^-(x) = -\frac{1}{2}\frac{g(x)}{X^+(x)} + \left(\mathcal{H}\frac{g}{X^+}\right)(x)\,. \tag{11.85c}$$

Substituting (11.85c) into (11.85a) and considering (11.76b) and $g(x) = f(x)/(a(x) + b(x))$ finally results in the the solution:

$$\varphi(x) = \frac{X^+(x) + X^-(x)}{2(a(x) + b(x))X^+(x)}\, f(x) + (X^+(x) - X^-(x))\,\frac{1}{2\pi\,\mathrm{i}} \int\limits_{\Gamma} \frac{f(y)}{(a(y) + b(y))X^+(y)(y-x)}\, dy + \varphi_h(x), \quad x \in \Gamma. \tag{11.86}$$

According to (11.83) in the case $\kappa < 0$ the following relations must hold simultaneously in order to ensure the existence of a solution:

$$\int\limits_{\Gamma} \frac{y^k f(y)}{(a(y) + b(y))X^+(y)}\, dy = 0 \quad (k = 0, 1, \ldots, -\kappa - 1). \tag{11.87}$$

■ The characteristic integral equation is given with constant coefficients a and b: $a\varphi(x) + \dfrac{b}{\pi\,\mathrm{i}}\displaystyle\int_{\Gamma} \frac{\varphi(y)}{y-x}\,dy = f(x)$. Here Γ is a simple closed curve, i.e., $\Gamma = \Gamma_0$ $(n = 0)$. From (11.77b) follows $G = \dfrac{a-b}{a+b}$ and $g(x) = \dfrac{f(x)}{a+b}$. G is a constant, consequently $\kappa = 0$. Therefore, $\Pi(x) = 1$ and $G_0 = G = \dfrac{a-b}{a+b}$. $\quad I(z) = \log\dfrac{a-b}{a+b}\dfrac{1}{2\pi\mathrm{i}}\displaystyle\int_{\Gamma}\frac{1}{y-z}dy = \begin{cases}\log\dfrac{a-b}{a+b}, & z \in S^+,\\ 0, & z \in S^-.\end{cases}$

$X(z) = \begin{cases}\dfrac{a-b}{a+b}, & z \in S^+,\\ 1, & z \in S^-,\end{cases}$ $\quad$ i.e., $\quad X^+ = \dfrac{a-b}{a+b},\ X^- = 1.$

Since $\kappa = 0$ holds, the homogeneous Hilbert boundary value problem has only the function $\Phi_h(z) = 0$ as the solution vanishing at infinity. From (11.86) follows

$$\varphi(x) = \frac{X^+ + X^-}{2(a+b)X^+}f(x) + \frac{X^+ - X^-}{2(a+b)X^+}\frac{1}{\pi\,\mathrm{i}}\int_{\Gamma}\frac{f(y)}{y-x}\,dy = \frac{a}{a^2-b^2}f(x) - \frac{b}{a^2-b^2}\frac{1}{\pi\,\mathrm{i}}\int_{\Gamma}\frac{f(y)}{y-x}dy.$$

12 Functional Analysis

1. Functional Analysis
Functional analysis arose after the recognition of a common structure in different disciplines such as the sciences, engineering and economics. General principles were discovered that resulted in a common and unified approach in calculus, linear algebra, geometry, and other mathematical fields, showing their interrelations.

2. Infinite Dimensional Spaces
There are many problems, the mathematical modeling of which requires the introduction of infinite systems of equations or inequalities. Differential or integral equations, approximation, variational or optimization problems could not be treated by using only finite dimensional spaces.

3. Linear and Non-Linear Operators
In the first phase of applying functional analysis – mainly in the first half of the twentieth century – linear or linearized problems were thoroughly examined, which resulted in the development of the theory of linear operators. More recently the application of functional analysis in practical problems required the development of the theory of non-linear operators, since more and more problems had to be solved that could be described only by non-linear methods. Functional analysis is increasingly used in solving differential equations, in numerical analysis and in optimization, and its principles and methods became a necessary tool in engineering and other applied sciences.

4. Basic Structures
In this chapter only the basic structures will be introduced, and only the most important types of abstract spaces and some special classes of operators in these spaces will be discussed. The abstract notion will be demonstrated by some examples, which are discussed in detail in other chapters of this book, and the existence and uniqueness theorems of the solutions of such problems are stated and proved there. Because of its abstract and general nature it is clear that functional analysis offers a large range of general relations in the form of mathematical theorems that can be directly used in solving a wide variety of practical problems.

12.1 Vector Spaces

12.1.1 Notion of a Vector Space

A non-empty set V is called a *vector space* or *linear space* over the field $\mathbb{F}$ of scalars if there exist two operations on V – addition of the elements and multiplication by scalars from $\mathbb{F}$ – such that they have the following properties:

1. for any two elements $x, y \in \mathrm{V}$, there exists an element $z = x + y \in \mathrm{V}$, which is called their *sum*.

2. For every $x \in \mathrm{V}$ and every scalar (number) $\alpha \in \mathbb{F}$ there exists an element $\alpha x \in \mathrm{V}$, the *product* of x and the scalar α so that the following properties, the *axioms of vector spaces* (see also 5.3.8.1, p. 365), are satisfied for arbitrary elements $x, y, z \in \mathrm{V}$ and scalars $\alpha, \beta \in \mathbb{F}$:

(V1) $x + (y + z) = (x + y) + z.$ (12.1)

(V2) There exists an element $0 \in \mathrm{V}$, the zero element, such that $x + 0 = x$. (12.2)

(V3) To every vector x there is a vector $-x$ such that $x + (-x) = 0$. (12.3)

(V4) $x + y = y + x.$ (12.4)

(V5) $1 \cdot x = x, \quad 0 \cdot x = 0.$ (12.5)

(V6) $\alpha(\beta x) = (\alpha\beta)x.$ (12.6)

(V7) $(\alpha + \beta)x = \alpha x + \beta x.$ (12.7)

(V8) $\alpha(x + y) = \alpha x + \alpha y.$ (12.8)

I.N. Bronshtein et al., *Handbook of Mathematics*,
DOI 10.1007/978-3-662-46221-8_12

V is called a real or complex vector space, depending on whether $\mathbb{F}$ is the field $\mathbb{R}$ of real numbers or the field $\mathbb{C}$ of complex numbers. The elements of V are also called either points or, according to linear algebra, *vectors*. The vector notation $\vec{x}$ or $\underline{x}$ is not used in functional analysis.

The difference $x - y$ of two arbitrary vectors $x, y \in$ V also can be defined in V as $x - y = x + (-y)$. From the previous definition, it follows that the equation $x + y = z$ can be solved uniquely for arbitrary elements y and z. The solution is $x = z - y$. Further properties follow from axioms **(V1)**–**(V8)**:

- the zero element is uniquely defined,
- $\alpha x = \beta x$ and $x \neq 0$, imply $\alpha = \beta$,
- $\alpha x = \alpha y$ and $\alpha \neq 0$, imply $x = y$,
- $-(\alpha x) = \alpha \cdot (-x)$.

12.1.2 Linear and Affine Linear Subsets

1. Linear Subsets

A non-empty subset V_0 of a vector space V is called a *linear subspace* or a *linear manifold* of V if together with two arbitrary elements $x, y \in \mathrm{V}_0$ and two arbitrary scalars $\alpha, \beta \in \mathbb{F}$, their linear combination $\alpha x + \beta y$ is also in V_0. V_0 is a vector space in its own right, and therefore satisfies the axioms **(V1)**–**(V8)**. The subspace V_0 can be V itself or only the zero point. In these cases the subspace is called trivial.

2. Affine Subspaces

A subset of a vector space V is called an *affine linear subspace* or an *affine manifold* if it has the form

$$\{x_0 + y : y \in \mathrm{V}_0\}, \tag{12.9}$$

where $x_0 \in$ V is a given element and V_0 is a linear subspace. It can be considered (in the case $x_0 \neq 0$) as the generalization of the lines or planes not passing through the origin in $\mathbb{R}^3$.

3. The Linear Hull

The intersection of an arbitrary number of subspaces in V is also a subspace. Consequently, for every non-empty subset $E \subset$ V, there exists a smallest linear subset $lin(E)$ or $[E]$ in V containing E, namely the intersection of all the linear subspaces, which contain E. The set $lin(E)$ is called the *linear hull of the set* E, or the *linear subspace generated by the set* E. It coincides with the set of all (finite) linear combinations

$$\alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_n x_n, \tag{12.10}$$

comprised of elements $x_1, x_2, \ldots, x_n \in E$ and scalars $\alpha_1, \alpha_2, \ldots, \alpha_n \in \mathbb{F}$.

4. Examples for Vector Spaces of Sequences

■ **A Vector Space $\mathbb{F}^n$:** Let n be a given natural number and V the set of all n-tuples, i.e., all finite sequences consisting of n scalar terms $\{(\xi_1, \ldots, \xi_n) : \ \xi_i \in \mathbb{F},\ i = 1, \ldots, n\}$. The operations will be defined componentwise or termwise, i.e., if $x = (\xi_1, \ldots, \xi_n)$ and $y = (\eta_1, \ldots, \eta_n)$ are two arbitrary elements from V and α is an arbitrary scalar, $\alpha \in \mathbb{F}$, then

$$x + y = (\xi_1 + \eta_1, \ldots, \xi_n + \eta_n), \ (12.11a) \qquad \alpha \cdot x = (\alpha\xi_1, \ldots, \alpha\xi_n). \ (12.11b)$$

In this way, the vector space $\mathbb{F}^n$ is defined. The linear spaces $\mathbb{R}$ or $\mathbb{C}$ are special cases for $n = 1$. This example can be generalized in two different ways (see examples **B** and **C**).

■ **B Vector Space s of all Sequences:** Considering the infinite sequences as elements $x = \{\xi_n\}_{n=1}^{\infty}$, $\xi_n \in \mathbb{F}$ and defining the operations componentwise, similar to (12.11a) and (12.11b), the vector space **s** of all sequences are obtained.

■ **C Vector Space φ(also $\mathbf{c_{00}}$) of all Finite Sequences:** Let V be the subset of all elements of **s** containing only a finite number of non-zero components, where the number of non-zero components depends on the element. This vector space – the operations are again introduced termwise – is denoted by $\boldsymbol{\varphi}$ or also by $\mathbf{c_{00}}$, and it is called the space of all *finite sequences of numbers*.

■ **D Vector Space m (also $\mathbf{l}^\infty$) of all Bounded Sequences:** A sequence $x = \{\xi_n\}_{n=1}^\infty$ belongs to **m** if and only if there exists $C_x > 0$ with $|\xi_n| \le C_x,\ \forall\, n = 1, 2, \dots$. This vector space is also denoted by $\mathbf{l}^\infty$.

■ **E Vector Space c of all Convergent Sequences:** A sequence $x = \{\xi_n\}_{n=1}^\infty$ belongs to **c** if and only if there exists a number $\xi_0 \in \mathbb{F}$ such that for $\forall\, \varepsilon > 0$ there exists an index $n_0 = n_0(\varepsilon)$ such that for all $n > n_0$ one has $|\xi_n - \xi_0| < \varepsilon$ (see 7.1.2, p. 458).

■ **F Vector Space $\mathbf{c_0}$ of all Null Sequences:** The vector space $\mathbf{c_0}$ of all null sequences, i.e., the subspace of **c** consisting of all sequences converging to zero ($\xi_0 = 0$).

■ **G Vector Space $\mathbf{l}^p$:** The vector space of all sequences $x = \{\xi_n\}_{n=1}^\infty$ such that $\sum_{n=1}^\infty |\xi_n|^p$ is convergent, is denoted by $\mathbf{l}^p$ $(1 \le p < \infty)$.
It can be shown by the Minkowski inequality that the sum of two sequences from $\mathbf{l}^p$ also belongs to $\mathbf{l}^p$, (see 1.4.2.13, p. 32).

Remark: For the vector spaces introduced in examples **A**–**G**, the following inclusions hold:

$$\varphi \subset \mathbf{c_0} \subset \mathbf{c} \subset \mathbf{m} \subset \mathbf{s} \quad \text{and} \quad \varphi \subset \mathbf{l}^p \subset \mathbf{l}^q \subset \mathbf{c_0}, \quad \text{where} \quad 1 \le p < q < \infty. \tag{12.12}$$

5. Examples of Vector Spaces of Functions

■ **A Vector Space $\mathcal{F}(T)$:** Let V be the set of all real or complex valued functions defined on a given set T, where the operations are defined point-wise, i.e., if $x = x(t)$ and $y = y(t)$ are two arbitrary elements of V and $\alpha \in \mathbb{F}$ is an arbitrary scalar, then we define the elements (functions) $x + y$ and $\alpha \cdot x$ by the rules

$$(x + y)(t) = x(t) + y(t) \quad \forall\, t \in T, \tag{12.13a}$$

$$(\alpha x)(t) = \alpha \cdot x(t) \quad \forall\, t \in T. \tag{12.13b}$$

This vector space is denoted by $\mathcal{F}(T)$.
Some of the subspaces are introduced in the following examples.

■ **B Vector Space $\mathcal{B}(T)$ or $\mathcal{M}(T)$:** The space $\mathcal{B}(T)$ is the space of all functions bounded on T. This vector space is often denoted by $\mathcal{M}(T)$. In the case of $T = \mathbb{N}$, one gets the space $\mathcal{M}(\mathbb{N}) = \mathbf{m}$ from example **D** of the previous paragraph.

■ **C Vector Space $\mathcal{C}([a, b])$:** The set $\mathcal{C}([a, b])$ of all functions continuous on the interval $[a, b]$ (see 2.1.5.1, p. 58).

■ **D Vector Space $\mathcal{C}^{(k)}([a, b])$:** Let $k \in \mathbb{N},\ k \ge 1$. The set $\mathcal{C}^{(k)}([a, b])$ of all functions k-times continuously differentiable on $[a, b]$ (see 6.1, p. 432–437) is a vector space. At the endpoints a and b of the interval $[a, b]$, the derivatives have to be considered as right-hand and left-hand derivatives, respectively.

Remark: For the vector spaces of examples **A**–**D** of this paragraph, and $T = [a, b]$ the following subspace relations hold:

$$\mathcal{C}^{(k)}([a, b]) \subset \mathcal{C}([a, b]) \subset \mathcal{B}([a, b]) \subset \mathcal{F}([a, b]). \tag{12.14}$$

■ **E Vector Subspace of $\mathcal{C}([a, b])$:** For any given point $t_0 \in [a, b]$, the set $\{x \in \mathcal{C}([a, b]) \colon x(t_0) = 0\}$ forms a linear subspace of $\mathcal{C}([a, b])$.

12.1.3 Linearly Independent Elements

1. Linear Independence

A finite subset $\{x_1, \dots, x_n\}$ of a vector space V is called *linearly independent* if

$$\alpha_1 x_1 + \cdots + \alpha_n x_n = 0 \quad \text{implies} \quad \alpha_1 = \cdots = \alpha_n = 0. \tag{12.15}$$

Otherwise, it is called *linearly dependent*. If $\alpha_1 = \cdots = \alpha_n = 0$, then for arbitrary vectors $x_1, \dots, x_n$ from V, the vector $\alpha_1 x_1 + \cdots + \alpha_n x_n$ is trivially the zero element of V. Linear independence of the vectors $x_1, \dots, x_n$ means that the only way to produce the zero element $0 = \alpha_1 x_1 + \cdots + \alpha_n x_n$ is when all coefficients are equal to zero $\alpha_1 = \cdots = \alpha_n = 0$. This important notion is well known from linear

algebra (see 5.3.8.2, p. 366) and was used e.g. for the definition of a fundamental system of solutions of linear homogeneous differential equations (see 9.1.2.3, **2.**, p. 553). An infinite subset $E \subset \mathrm{V}$ is called *linearly independent* if every finite subset of E is linearly independent. Otherwise, E is called *linearly dependent*.

■ If the sequence whose k-th term is equal to 1 and all the others are 0 is denoted by e_k , then belongs e_k to the space $\boldsymbol{\varphi}$ and consequently to any space of sequences. The set $\{e_1, e_2, \dots\}$ is linearly independent in every one of these spaces. In the space $\mathcal{C}([0,\pi])$, e.g., the system of functions

$$1,\ \sin nt,\ \cos nt \quad (n = 1, 2, 3, \dots)$$

is linearly independent, but the functions $1, \cos 2t, \cos^2 t$ are linearly dependent (see (2.97), p. 81).

2. Basis and Dimension of a Vector Space

A linearly independent subset B from V, which generates the whole space V, i.e., $lin(B) = \mathrm{V}$ holds, is called an *algebraic basis* or a *Hamel basis* of the vector space V (see 5.3.8.2, p. 366). $B = \{x_\xi : \xi \in \Xi\}$ is a basis of V if and only if every vector $x \in \mathrm{V}$ can be written in the form $x = \sum\limits_{\xi\in\Xi} \alpha_\xi x_\xi$, where the coefficients α_ξ are uniquely determined by x and only a finite number of them (depending on x) can be different from zero. Every non-trivial vector space V, i.e., $\mathrm{V} \neq \{0\}$, has at least one algebraic basis, and for every linearly independent subset E of V, there exists at least one algebraic basis of V, which contains this subset of E.

A vector space V is *m-dimensional* if it possesses a basis consisting of m vectors. That is, there exist m linearly independent vectors in V, and every system of $m + 1$ vectors is linearly dependent.

A vector space is *infinite dimensional* if it has no finite basis, i.e., if for every natural number m there are m linearly independent vectors in V.

The space $\mathbb{F}^n$ is n-dimensional, and all the other spaces in examples **B**–**E** are infinite dimensional. The subspace $lin(\{1, t, t^2\}) \subset \mathcal{C}([a, b])$ is three-dimensional.

In the finite dimensional case, every two bases of the same vector space have the same number of elements. Also in an infinite dimensional vector space any two bases have the same cardinality, which is denoted by $dim(\mathrm{V})$. The dimension is an invariant quantity of the vector space, it does not depend on the particular choice of an algebraic basis.

12.1.4 Convex Subsets and the Convex Hull

12.1.4.1 Convex Sets

A subset C of a real vector space V is called *convex* if for every pair of vectors $x, y \in C$ all vectors of the form $\lambda x + (1 - \lambda)y,\ 0 \le \lambda \le 1$, also belong to C. In other words, the set C is convex, if for any two elements x and y, the whole line segment

$$\{\lambda x + (1 - \lambda)y :\ 0 \le \lambda \le 1\}, \tag{12.16}$$

(which is also called an interval), belongs to C. (For examples of convex sets in $\mathbb{R}^2$ see the sets denoted by A and B in **Fig. 12.5**, p. 684.)

The intersection of an arbitrary number of convex sets is also a convex set, where the empty set is agreed to be convex. Consequently, for every subset $E \subset \mathrm{V}$ there exists a smallest convex set which contains E, namely, the intersection of all convex subsets of V containing E. It is called the *convex hull* of the set E and it is denoted by $co\,(E)$. $co\,(E)$ is identical to the set of all finite *convex* linear combinations of elements from E, i.e., $co\,(E)$ consists of all elements of the form $\lambda_1 x_1 + \dots + \lambda_n x_n$, where $x_1, \dots, x_n$ are arbitrary elements from E and $\lambda_i \in [0, 1]$ satisfy the equality $\lambda_1 + \dots + \lambda_n = 1$. Linear and affine subspaces are always convex.

12.1.4.2 Cones

A non-empty subset C of a (real) vector space V is called a *convex cone* if it satisfies the following properties:

1. C is a convex set.

2. From $x \in C$ and $\lambda \geq 0$, it follows that $\lambda x \in C$.

3. From $x \in C$ and $-x \in C$, it follows that $x = 0$.

A cone can be characterized also by **3.** together with

$$x, y \in C \quad \text{and} \quad \lambda, \mu \geq 0 \quad \text{imply} \quad \lambda x + \mu y \in C. \tag{12.17}$$

■ **A:** The set $\mathbb{R}^n_+$ of all vectors $x = (\xi_1, \dots, \xi_n)$ with non-negative components is a cone in $\mathbb{R}^n$.

■ **B:** The set $\mathcal{C}_+$ of all real continuous functions on $[a, b]$ with only non-negative values is a cone in the space $\mathcal{C}([a, b])$.

■ **C:** The set of all sequences of real numbers $\{\xi_n\}_{n=1}^{\infty}$ with only non-negative terms, i.e., $\xi_n \geq 0,\ \forall n$, is a cone in **s**. Analogously, cones are obtained in the spaces of examples **C**–**G** in 12.1.2, p. 655, if the sets of non-negative sequences are considered in these spaces.

■ **D:** The set $C \subset \mathbf{l}^p$ $(1 \leq p < \infty)$, consisting of all sequences $\{\xi_n\}_{n=1}^{\infty}$, such that for some $a > 0$

$$\sum_{n=1}^{\infty} |\xi_n|^p \leq a \tag{12.18}$$

is a convex set in $\mathbf{l}^p$, but obviously, not a cone.

■ **E:** Examples from $\mathbb{R}^2$ see **Fig. 12.1**: **a)** convex set, not a cone, **b)** not convex, **c)** convex hull.

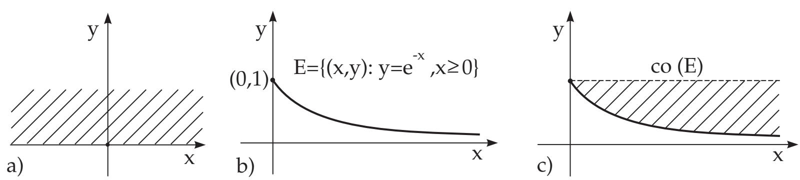


Figure 12.1

12.1.5 Linear Operators and Functionals

12.1.5.1 Mappings

A mapping $T\colon D \longrightarrow \mathbb{Y}$ from the set $D \subset \mathbb{X}$ into the set $\mathbb{Y}$ is called

• *injective,* if

$$T(x) = T(y) \Longrightarrow x = y, \tag{12.19}$$

• *surjective,* if for

$$\forall\, y \in \mathbb{Y} \quad \text{there exists an element} \quad x \in D \quad \text{such that} \quad T(x) = y, \tag{12.20}$$

• *bijective,* if T is both injective and surjective.

D is called the *domain* of the mapping T and is denoted by D_T or $D(T)$, while the subset $\{y \in \mathbb{Y} : \exists\, x \in D_T \text{ with } T(x) = y\}$ of $\mathbb{Y}$ is called the *range* of the mapping T and is denoted by $\mathcal{R}(T)$ or $Im(T)$.

12.1.5.2 Homomorphism and Endomorphism

Let $\mathbb{X}$ and $\mathbb{Y}$ be two vector spaces over the same field $\mathbb{F}$ and D a linear subset of $\mathbb{X}$. A mapping $T\colon D \longrightarrow \mathbb{Y}$ is called *linear* (or a *linear transformation, linear operator* or *homomorphism*), if for arbitrary $x, y \in D$ and $\alpha, \beta \in \mathbb{F}$,

$$T(\alpha x + \beta y) = \alpha Tx + \beta Ty. \tag{12.21}$$

For a linear operator T the notation Tx is preferred, which is similarly used for linear functions, while the notation $T(x)$ is used for general operators.

The range $\mathcal{R}(T)$ is the set of all $y \in \mathrm{Y}$ such that the equation $Tx = y$ has at least one solution. $N(T) = \{x \in \mathrm{X} : Tx = 0\}$ is the *null space* or *kernel* of the operator T and is also denoted by $\ker(T)$.

A mapping of the vector space X into itself is called an *endomorphism*. If T is an injective linear mapping, then the mapping defined on $\mathcal{R}(T)$ by

$$y \longmapsto x, \text{ such that } Tx = y,\ y \in \mathcal{R}(T) \tag{12.22}$$

is linear. It is denoted by T^{-1}: $\mathcal{R}(T) \longrightarrow \mathrm{X}$ and is called the *inverse* of T. If Y is the vector space $\mathbb{F}$, then a linear mapping f: $\mathrm{X} \longrightarrow \mathbb{F}$ is called a *linear functional* or a *linear form*.

12.1.5.3 Isomorphic Vector Spaces

A bijective linear mapping T: $\mathrm{X} \longrightarrow \mathrm{Y}$ is called an *isomorphism* of the vector spaces X and Y. Two vector spaces are called *isomorphic* provided an isomorphism exists.

12.1.6 Complexification of Real Vector Spaces

Every real vector space V can be extended to a complex vector space $\tilde{\mathrm{V}}$. The set $\tilde{\mathrm{V}}$ consists of all pairs (x, y) with $x, y \in \mathrm{V}$. The operations (addition and multiplication by a complex number $a + \mathrm{i}b \in \mathbb{C}$) are defined as follows:

$$(x_1, y_1) + (x_2, y_2) = (x_1 + x_2, y_1 + y_2), \tag{12.23a}$$
$$(a+\mathrm{i}b)(x, y) = (ax - by, bx + ay). \tag{12.23b}$$

Since the special relations

$$(x, y) = (x, 0) + (0, y) \text{ and } \mathrm{i}(y, 0) = (0 + \mathrm{i}1)(y, 0) = (0 \cdot y - 1 \cdot 0, 1y + 0 \cdot 0) = (0, y) \tag{12.24}$$

hold, the pair (x, y) can also be written as $x + \mathrm{i}y$. The set $\tilde{\mathrm{V}}$ is a complex vector space, where the set V is identified with the linear subspace $\tilde{\mathrm{V}}_0 = \{(x, 0) \colon x \in \mathrm{V}\}$, i.e., $x \in \mathrm{V}$ is considered as $(x, 0)$ or as $x + \mathrm{i}0$.

This procedure is called the *complexification* of the vector space V. A linearly independent subset in V is also linearly independent in $\tilde{\mathrm{V}}$. The same statement is valid for a basis in V, so $\mathit{dim}(\mathrm{V}) = \mathit{dim}(\tilde{\mathrm{V}})$.

12.1.7 Ordered Vector Spaces

12.1.7.1 Cone and Partial Ordering

If a cone C is fixed in a vector space V, then an *order* can be introduced for certain pairs of vectors in V. Namely, if $x - y \in C$ for some $x, y \in \mathrm{V}$ then one writes $x \geq y$ or $y \leq x$ and say x is *greater than or equal to* y or y is *smaller than or equal to* x. The pair (V, C) is called an *ordered vector space* or a vector space *partially ordered by the cone* C. An element x is called *positive*, if $x \geq 0$ or, which means the same, if $x \in C$ holds. Moreover

$$C = \{x \in \mathrm{V} \colon x \geq 0\}. \tag{12.25}$$

If the vector space $\mathbb{R}^2$ ordered by its first quadrant as the cone $C(= \mathbb{R}^2_+)$ is under consideration, then a typical phenomenon of ordered vector spaces will be seen. This is referred to as "partially ordered" or sometimes as "semi-ordered". Namely, only certain pairs of two vectors are comparable. Considering the vectors $x = (1, -1)$ and $y = (0, 2)$, neither the vector $x - y = (1, -3)$ nor $y - x = (-1, 3)$ is in C, so neither $x \geq y$ nor $x \leq y$ holds. An ordering in a vector space, generated by a cone, is always only a partial ordering.

It can be shown that the binary relation $\geq$ has the following properties:

(O1) $x \geq x \ \forall x \in \mathrm{V}$ (reflexivity). (12.26)

(O2) $x \geq y$ and $y \geq z$ imply $x \geq z$ (transitivity). (12.27)

(O3) $x \geq y$ and $\alpha \geq 0$, $\alpha \in \mathbb{R}$, imply $\alpha x \geq \alpha y$. (12.28)

(O4) $x_1 \geq y_1$ and $x_2 \geq y_2$ imply $x_1 + x_2 \geq y_1 + y_2$. (12.29)

Conversely, if in a vector space V there exists an ordering relation, i.e., a binary relation $\geq$ is defined for certain pairs of elements and satisfies axioms **(O1)**–**(O4)**, and if one puts

$$\mathrm{V}_+ = \{x \in \mathrm{V}\colon x \geq 0\}, \tag{12.30}$$

then it can be shown that V_+ is a cone. The order $\geq_{V_+}$ in V induced by V_+ is identical to the original order $\geq$; consequently, the two possibilities of introducing an order in a vector space are equivalent. A cone $C \subset \mathrm{V}$ is called *generating* or *reproducing* if every element $x \in \mathrm{V}$ can be represented as $x = u - v$ with $u, v \in C$. It can be written in the form $\mathrm{V} = C–C$.

■ **A:** An obvious order in the space **s** (see example **B**, p. 655) is induced by means of the cone

$$C = \{x = \{\xi_n\}_{n=1}^{\infty} : \xi_n \geq 0 \quad \forall\, n\} \tag{12.31}$$

(see example **C**, p. 658).

In the spaces of sequences (see (12.12), p 656) usually the natural coordinate-wise order is considered. This is defined by the cone obtained as the intersection of the considered space with C (see (12.31), p. 660). The positive elements in these ordered vector spaces are then the sequences with non-negative terms. It is clear that other orders can be defined by other cones, as well. Then orderings different from the natural ordering can be obtained (see [12.17], [12.19]).

■ **B:** In the real spaces of functions $\mathcal{F}(T),\ \mathcal{B}(T),\ \mathcal{C}([a,b])$ and $\mathcal{C}^{(k)}([a,b])$ (see 12.1.2, **5.**, p. 656), the natural order $x \geq y$ for two functions x and y is defined by $x(t) \geq y(t)$, $\forall\, t \in T$, or $\forall\, t \in [a,b]$. Then $x \geq 0$ if and only if x is a non-negative function in T. The corresponding cones are denoted by $\mathcal{F}_+(T),\ \mathcal{B}_+(T)$, etc. Also $C_+ = \mathcal{C}_+(T) = \mathcal{F}_+(T) \cap \mathcal{C}(T)$ can be obtained if $T = [a,b]$.

12.1.7.2 Order Bounded Sets

Let E be an arbitrary non-empty subset of an ordered vector space V. An element $z \in \mathrm{V}$ is called an *upper bound* of the set E if for every $x \in E$, $x \leq z$. An element $u \in \mathrm{V}$ is a *lower bound* of E if $u \leq x$, $\forall\, x \in E$. For any two elements $x, y \in \mathrm{V}$ with $x \leq y$, the set

$$[x,y] = \{v \in \mathrm{V}\colon x \leq v \leq y\} \tag{12.32}$$

is called an *order interval* or (o)-*interval.*

Obviously, the elements x and y are a lower bound and an upper bound of the set $[x,y]$, respectively, where they even belong to the set. A set $E \subset \mathrm{V}$ is called *order bounded* or simply (o) *bounded*, if E is a subset of an order interval, i.e., if there exist two elements $u, z \in \mathrm{V}$ such that $u \leq x \leq z$, $\forall\, x \in E$ or, equivalently, $E \subset [u,z]$. A set is called *bounded above* or *bounded below* if it has an upper bound, or a lower bound, respectively.

12.1.7.3 Positive Operators

A linear operator (see [12.2], [12.17]) $T\colon \mathrm{X} \longrightarrow \mathrm{Y}$ from an ordered vector space $\mathrm{X} = (\mathrm{X}, \mathrm{X}_+)$ into an ordered vector space $\mathrm{Y} = (\mathrm{Y}, \mathrm{Y}_+)$ is called *positive*, if

$$T(\mathrm{X}_+) \subset \mathrm{Y}_+, \quad \text{i.e.,} \quad Tx \geq 0 \quad \text{for all} \quad x \geq 0. \tag{12.33}$$

12.1.7.4 Vector Lattices

1. Vector Lattices

In the vector space $\mathbb{R}^1$ of the real numbers the notions of (o)-boundedness and boundedness (in the usual sense) are identical. It is known that every set of real numbers which is bounded from above has a supremum: the smallest of its upper bounds (or the least upper bound, sometimes denoted by lub). Analogously, if a set of reals is bounded from below, then it has an *infimum*, the greatest lower bound, sometimes denoted by glb. In a general ordered vector space, the existence of the supremum and infimum cannot be guaranteed even for finite sets. They must be given by axioms. An ordered vector space V is called a *vector lattice* or a *linear lattice* or a *Riesz space*, if for two arbitrary elements $x, y \in \mathrm{V}$ there exists an element $z \in \mathrm{V}$ with the following properties:

1. $x \leq z$ and $y \leq z$,

2. if $u \in \mathrm{V}$ with $x \leq u$ and $y \leq u$, then $z \leq u$.

Such an element z is uniquely determined, it is denoted by $x \vee y$, and it is called the *supremum* of x and y (more precisely: supremum of the set consisting of the elements x and y). In a vector lattice, there also exists the infimum for any x and y, which is denoted by $x \wedge y$. For applications of positive operators in vector lattices see, e.g., [12.2], [12.3] [12.15].

A vector lattice is called *Dedekind complete* or a *K-space* (*Kantorovich space*) if every non-empty subset E that is order bounded from above has a supremum lub(E) (equivalently, if every non-empty subset that is order bounded from below has an infimum glb(E)).

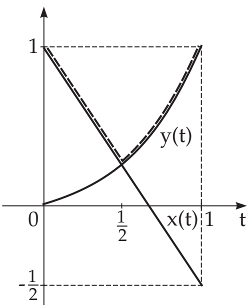


Figure 12.2

■ **A:** In the vector lattice $\mathcal{F}([a,b])$ (see 12.1.2, **5.**, p. 656), the supremum of two functions x, y is calculated pointwise by the formula

$$(x \vee y)(t) = \max\{x(t), y(t)\} \quad \forall\, t \in [a,b]. \tag{12.34}$$

In the case of $[a,b] = [0,1]$, $x(t) = 1 - \frac{3}{2}t$ and $y(t) = t^2$ **(Fig. 12.2)**,

$$(x \vee y)(t) = \begin{cases} 1 - \frac{3}{2}t, & \text{if} \quad 0 \le t \le \frac{1}{2}\,, \\ t^2, & \text{if} \quad \frac{1}{2} \le t \le 1 \end{cases} \tag{12.35}$$

is obtained.

■ **B:** The spaces $\mathcal{C}([a,b])$ and $\mathcal{B}([a,b])$ (see 12.1.2, **5.**, p. 656) are also vector lattices, while the ordered vector space $\mathcal{C}^{(1)}([a,b])$ is not a vector lattice, since the minimum or maximum of two differentiable functions may not be differentiable on $[a,b]$, in general.

A linear operator $T\colon\ \mathrm{X} \longrightarrow \mathrm{Y}$ from a vector lattice X into a vector lattice Y is called a *vector lattice homomorphism* or *homomorphism of the vector lattice*, if for all $x, y \in \mathrm{X}$

$$T(x \vee y) = Tx \vee Ty \quad \text{and} \quad T(x \wedge y) = Tx \wedge Ty. \tag{12.36}$$

2. Positive and Negative Parts, Modulus of an Element

For an arbitrary element x of a vector lattice V, the elements

$$x_+ = x \vee 0, \quad x_- = (-x) \vee 0 \quad \text{and} \quad |x| = x_+ + x_- \tag{12.37}$$

are called the *positive part, negative part*, and *modulus* of the element x, respectively. For every element $x \in \mathrm{V}$, the three elements $x_+, x_-, |x|$ are positive, where for $x, y \in \mathrm{V}$ the following relations are valid:

$$x \le x_+ \le |x|, \quad x = x_+ - x_-, \quad x_+ \wedge x_- = 0, \quad |x| = x \vee (-x), \tag{12.38a}$$

$$(x+y)_+ \le x_+ + y_+, \quad (x+y)_- \le x_- + y_-, \quad |x+y| \le |x| + |y|, \tag{12.38b}$$

$$x \le y \quad \text{implies} \quad x_+ \le y_+ \quad \text{and} \quad x_- \ge y_- \tag{12.38c}$$

and for arbitrary $\alpha \ge 0$

$$(\alpha\, x)_+ = \alpha\, x_+, \quad (\alpha\, x)_- = \alpha\, x_-, \quad |\alpha\, x| = \alpha |x|. \tag{12.38d}$$

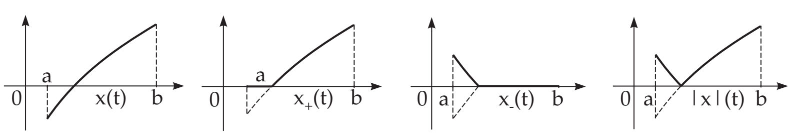


Figure 12.3

In the vector spaces $\mathcal{F}([a,b])$ and $\mathcal{C}([a,b])$, the positive part, the negative part, and the modulus of a function $x(t)$ can be got by means of the following formulas **(Fig. 12.3)**:

$$x_+(t) = \begin{cases} x(t), & \text{if } x(t) \ge 0, \\ 0, & \text{if } x(t) < 0, \end{cases} \tag{12.39a}$$

$$x_-(t) = \begin{cases} 0, & \text{if } x(t) > 0, \\ -x(t), & \text{if } x(t) \le 0, \end{cases} \tag{12.39b}$$ $$|x|(t) = |x(t)| \quad \forall\, t \in [a,b]. \tag{12.39c}$$

12.2 Metric Spaces

12.2.1 Notion of a Metric Space

Let X be a set, and suppose a real, non-negative function $\rho(x,y)$ $(x, y \in \mathrm{X})$ is defined on $\mathrm{X} \times \mathrm{X}$. If this function $\rho:\ \mathrm{X} \times \mathrm{X} \to \mathbb{R}^1_+$ satisfies the following properties **(M1)**–**(M3)** for arbitrary elements $x, y, z \in \mathrm{X}$, then it is called a *metric* or *distance* in the set X, and the pair $\mathrm{X} = (\mathrm{X}, \rho)$ is called a *metric space*. The axioms of *metric spaces* are:

(M1) $\rho(x,y) \ge 0$ and $\rho(x,y) = 0$ if and only if $x = y$ (non-negativity), (12.40)

(M2) $\rho(x,y) = \rho(y,x)$ (symmetry), (12.41)

(M3) $\rho(x,y) \le \rho(x,z) + \rho(z,y)$ (triangle inequality). (12.42)

A metric can be defined on every subset Y of a metric space $\mathrm{X} = (\mathrm{X}, \rho)$ in a natural way if the metric ρ of the space X is restricted to the set Y, i.e., if ρ is considered only on the subset $\mathrm{Y} \times \mathrm{Y}$. The space (Y, ρ) of $\mathrm{X} \times \mathrm{X}$ is called a *subspace* of the metric space X.

■ **A:** The sets $\mathbb{R}^n$ and $\mathbb{C}^n$ are metric spaces with the *Euclidean metric* defined for points $x = (\xi_1, \ldots, \xi_n)$ and $y = (\eta_1, \ldots, \eta_n)$ as

$$\rho(x,y) = \sqrt{\sum_{k=1}^{n} |\xi_k - \eta_k|^2}. \tag{12.43}$$

■ **B:** The function

$$\rho(x,y) = \max_{1 \le k \le n} |\xi_k - \eta_k| \tag{12.44}$$

for vectors $x = (\xi_1, \ldots, \xi_n)$ and $y = (\eta_1, \ldots, \eta_n)$ also defines a metric in $\mathbb{R}^n$ and $\mathbb{C}^n$, the so-called *maximum metric*. If $\tilde{x} = (\tilde{\xi}_1, \ldots, \tilde{\xi}_n)$ is an approximation of the vector x, then it is of interest to know how much is the maximal deviation between the coordinates: $\max\limits_{1 \le k \le n} |\xi_k - \tilde{\xi}_k|$.

The function

$$\rho(x,y) = \sum_{k=1}^{n} |\xi_k - \eta_k| \tag{12.45}$$

for vectors $x, y \in \mathbb{R}^n$ (or $\mathbb{C}^n$) defines a metric in $\mathbb{R}^n$ and $\mathbb{C}^n$, the so-called *absolute value metric*. The metrics (12.43), (12.44) and (12.45) are reduced in the case of $n = 1$ to the absolute value $|x - y|$ in the spaces $\mathbb{R}$ and $\mathbb{C}$ (the sets of real and complex numbers).

■ **C:** Finite 0-1 sequences, e.g., 1110 and 010110, are called *words* in coding theory. If the number of positions is counted where two words of the same length n have different digits, i.e., for $x = (\xi_1, \ldots, \xi_n)$, $y = (\eta_1, \ldots, \eta_n)$, $\xi_k, \eta_k \in \{0, 1\}$, $\varrho(x,y)$ is defined as the number of the $k \in \{1, \ldots, n\}$ values such that $\xi_k \ne \eta_k$, then the set of words with a given length n is a metric space, and the metric is the so-called *Hamming distance*, e.g., $\varrho((1110), (0100)) = 2$.

■ **D:** In the set **m** and in its subsets **c** and $\mathbf{c_0}$ (see (12.12), p. 656) a metric is defined by

$$\rho(x,y) = \sup_k |\xi_k - \eta_k|, \quad (x = (\xi_1, \xi_2, \ldots), y = \eta_1, \eta_2, \ldots)). \tag{12.46}$$

■ **E:** In the set $\mathbf{l}^p$ $(1 \le p < \infty)$ of sequences $x = (\xi_1, \xi_2, \ldots)$ with absolutely convergent series $\sum\limits_{n=1}^{\infty} |\xi_n|^p$ a metric is defined by

$$\rho(x, y) = \sqrt[p]{\sum_{n=1}^{\infty} |\xi_n - \eta_n|^p}, \quad (x, y \in l^p). \tag{12.47}$$

■ **F:** In the set $\mathcal{C}([a, b])$ a metric is defined by

$$\rho(x, y) = \max_{t \in [a,b]} |x(t) - y(t)|. \tag{12.48}$$

■ **G:** In the set $\mathcal{C}^{(k)}([a, b])$ a metric is defined by

$$\rho(x, y) = \sum_{l=0}^{k} \max_{t \in [a,b]} |x^{(l)}(t) - y^{(l)}(t)|, \tag{12.49}$$

where (see (12.14) $\mathcal{C}^{(0)}([a, b])$ is understood as $\mathcal{C}([a, b])$).

■ **H:** Consider the set $L^p(\Omega)$ $(1 \le p < \infty)$ of the equivalence classes of Lebesgue measurable functions which are defined almost everywhere on a bounded domain $\Omega \subset \mathbb{R}^n$ and $\int\limits_{\Omega} |x(t)|^p \, d\mu < \infty$ (see also 12.9, p. 693). A metric in this set is defined by

$$\rho(x, y) = \sqrt[p]{\int_{\Omega} |x(t) - y(t)|^p \, d\mu}. \tag{12.50}$$

12.2.1.1 Balls, Neighborhoods and Open Sets

In a metric space $\mathrm{X} = (\mathrm{X}, \rho)$, whose elements are also called points, the following sets

$$B(x_0; r) = \{x \in \mathrm{X} : \rho(x, x_0) < r\}, \ (12.51) \qquad \overline{B}(x_0; r) = \{x \in \mathrm{X} : \rho(x, x_0) \le r\} \tag{12.52}$$

defined by means of a real number $r > 0$ and a fixed point x_0, are called an *open* and *closed ball* with radius r and center at x_0, respectively.

The balls (circles) defined by the metrics (12.43) and (12.44) and (12.45) in the vector space $\mathbb{R}^2$ are represented in **Fig. 12.4a,b** with $x_0 = 0$ and $r = 1$.

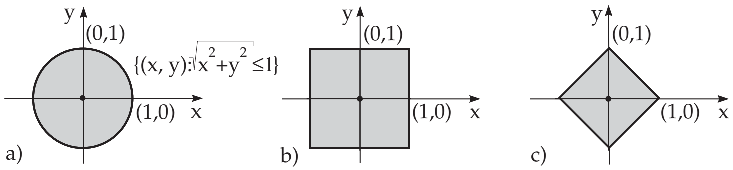


Figure 12.4

A subset U of a metric space $\mathrm{X} = (\mathrm{X}, \rho)$ is called a *neighborhood* of the point x_0 if U contains x_0 together with an open ball centered at x_0, in other words, if there exists an $r > 0$ such that $B(x_0; r) \subset U$. A neighborhood U of the point x is also denoted by $U(x)$. Obviously, every ball is a neighborhood of its center; an open ball is a neighborhood of all of its points. A point x_0 is called an *interior point* of a set $A \subset \mathrm{X}$ if x_0 belongs to A together with some of its neighborhood, i.e., there is a neighborhood U of x_0 such that $x_0 \in U \subset A$. A subset of a metric space is called *open* if all of its points are interior points. Obviously, X is an open set.

The open balls in every metric space, especially the open intervals in $\mathbb{R}$, are the prototypes of open sets.

The set of all open sets satisfies the following *axioms of open sets*:

- If G_α is open for $\forall\alpha \in I$, then the set $\bigcup\limits_{\alpha\in I} G_\alpha$ is also open.
- If $G_1, G_2, \dots, G_n$ are finitely many arbitrary open sets, then the set $\bigcap\limits_{k=1}^{n} G_k$ is also open.
- The empty set $\emptyset$ is open by definition.

A subset A of a metric space is *bounded* if for a certain element x_0 (which does not necessarily belong to A) and a real number $R > 0$ the set A is in the ball $B(x_0; R)$, i.e., $\rho(x, x_0) < R$ for all $x \in A$.

12.2.1.2 Convergence of Sequences in Metric Spaces

Let $\mathrm{X} = (\mathrm{X}, \rho)$ be a metric space, $x_0 \in \mathrm{X}$ and $\{x_n\}_{n=1}^\infty,\ x_n \in \mathrm{X}$ a sequence of elements of X.
The sequence $\{x_n\}_{n=1}^\infty$ is called *convergent* to the point x_0 if for every neighborhood $U(x_0)$ there is an index $n_0 = n_0(U(x_0))$ such that for all $n > n_0$, $x_n \in U(x_0)$. The usual notation

$$x_n \longrightarrow x_0 \quad (n \to \infty) \quad \text{or} \quad \lim_{n\to\infty} x_n = x_0 \tag{12.53}$$

is used and the point x_0 is called the *limit of the sequence* $\{x_n\}_{n=1}^\infty$. The limit of a sequence is uniquely determined. Instead of an arbitrary neighborhood of the point x_0, it is sufficient to consider only open balls with arbitrary radii, so (12.53) is equivalent to the following: $\forall\, \varepsilon > 0$ (now thinking about the open ball $B(x_0; \varepsilon)$), there is an index $n_0 = n_0(\varepsilon)$, such that if $n > n_0$, then $\rho(x_n, x_0) < \varepsilon$. Notice that (12.53) means $\rho(x_n, x_0) \longrightarrow 0$.

With these notions introduced in special metric spaces the distance between points can be calculated and the convergence of point sequences can be investigated. This has a great importance in numerical methods and in approximating functions by certain classes of functions (see, e.g., 19.6, p. 982).
In the space $\mathbb{R}^n$, equipped with one of the metrics given above, convergence always means coordinate-wise convergence.
In the spaces $\mathcal{B}([a, b])$ and $\mathcal{C}([a, b])$, the convergence introduced by (12.48) means uniform convergence of the function sequence on the set $[a, b]$ (see 7.3.2, p. 468).
In the space $L^2(\Omega)$ convergence with respect to the metric (12.50) means convergence in the (quadratic) mean, i.e., $x_n \to x_0$ if

$$\int\limits_\Omega |x_n - x_0|^2\, d\mu \longrightarrow 0 \quad \text{for} \quad n \to \infty\,. \tag{12.54}$$

12.2.1.3 Closed Sets and Closure

1. Closed Sets

A subset F of a metric space X is called *closed* if $\mathrm{X} \setminus F$ is an open set. Every closed ball in a metric space, especially every interval of the form $[a, b],\ [a, \infty), (-\infty, a]$ in $\mathbb{R}$, is a closed set.
Corresponding to the axioms of open sets, the collection of all closed sets of a metric space has the following properties:

- If F_α are closed for $\forall\alpha \in I$, then the set $\bigcap\limits_{\alpha\in I} F_\alpha$ is closed.
- If $F_1, \dots, F_n$ are finitely many closed sets, then the set $\bigcup\limits_{k=1}^{n} F_k$ is closed.
- The empty set $\emptyset$ is a closed set by definition.

The sets $\emptyset$ and X are open and closed at the same time.
A point x_0 of a metric space X is called a *limit point* of the subset $A \subset \mathrm{X}$ if for every neighborhood $U(x_0)$,

$$U(x_0) \cap A \neq \emptyset. \tag{12.55}$$

If this intersection always contains at least one point different from x_0, then x_0 is called an *accumulation point* of the set A. A limit point, which is not an accumulation point, is called an *isolated point*.
An accumulation point of A does not need to belong to the set A, e.g., the point a with respect to the set $A = (a, b]$, while an isolated point of A must belong to the set A.

A point x_0 is a limit point of the set A if there exists a sequence $\{x_n\}_{n=1}^{\infty}$ with elements x_n from A, which converges to x_0. If x_0 is an isolated point, then $x_n = x_0,\ \forall\, n \geq n_0$ for some index n_0.

2. The Closure of a Set

Every subset A of a metric space X obviously lies in the closed set X. Therefore, there always exists a smallest closed set containing A, namely the intersection of all closed sets of X, which contain A. This set is called the *closure* of the set A and it is usually denoted by $\overline{A}$. $\overline{A}$ is identical to the set of all limit points of A; $\overline{A}$ is obtained from the set A by adding all of its accumulation points to it. A is a closed set if and only if $A = \overline{A}$. Consequently, closed sets can be characterized by sequences in the following way: A is closed if and only if for every sequence $\{x_n\}_{n=1}^{\infty}$ of elements of A, which converges in X to an element $x_0 (\in \mathrm{X})$, the limit x_0 also belongs to A.

Boundary points of A are defined as follows: x_0 is a boundary point of A if for every neighborhood $U(x_0)$, $U(x_0) \cap A \neq \emptyset$ and also $U(x_0) \cap (X \setminus A) \neq \emptyset$. x_0 itself does not need to belong to A. Another characterization of a closed set is the following: A is closed if it contains all of its boundary points. (The set of boundary points of the metric space X is the empty set.)

12.2.1.4 Dense Subsets and Separable Metric Spaces

A subset A of a metric space X is called *everywhere dense* if $\overline{A} = \mathrm{X}$, i.e., each point $x \in \mathrm{X}$ is a limit point of the set A. That is, for each $x \in \mathrm{X}$, there is a sequence $\{x_n\}\ \ x_n \in A$ such that $x_n \longrightarrow x$.

■ **A:** According to the Weierstrass approximation theorem, every continuous function on a bounded closed interval $[a, b]$ can be approximated arbitrarily well by polynomials in the metric space of the space $\mathcal{C}([a, b])$, i.e., uniformly. This theorem can now be formulated as follows: The set of polynomials on the interval $[a, b]$ is everywhere dense in $\mathcal{C}([a, b])$.

■ **B:** Further examples for everywhere dense subsets are the set of rational numbers $\mathbf{Q}$ and the set of irrational numbers in the space of the real numbers $\mathbb{R}$.

A metric space X is called *separable* if there exists a countable everywhere dense subset in X. A countable everywhere dense subset in $\mathbb{R}^n$ is, e.g., the set of all vectors with rational components. The space $\mathbf{l} = \mathbf{l}^1$ is also separable, since a countable everywhere dense subset is formed, for example, by the set of its elements of the form $x = (r_1, r_2, \ldots, r_N, 0, 0, \ldots)$, where r_i are rational numbers and $N = N(x)$ is an arbitrary natural number. The space $\mathbf{m}$ is not separable.

12.2.2 Complete Metric Spaces

12.2.2.1 Cauchy Sequences

Let $\mathrm{X} = (\mathrm{X}, \rho)$ be a metric space. A sequence $\{x_n\}_{n=1}^{\infty}$ with $x_n \in \mathrm{X}$ is called a *Cauchy sequence* if for $\forall\, \varepsilon > 0$ there is an index $n_0 = n_0(\varepsilon)$ such that for $\forall\, n, m > n_0$ there holds the inequality

$$\rho(x_n, x_m) < \varepsilon. \tag{12.56}$$

Every Cauchy sequence is a bounded set. Furthermore, every convergent sequence is a Cauchy sequence. In general, the converse statement is not true, as is shown in the following example.

■ Consider the space $\mathbf{l}^1$ with the metric (12.46) of the space $\mathbf{m}$. Obviously, the elements $x^{(n)} = \left(1, \frac{1}{2}, \frac{1}{3}, \ldots, \frac{1}{n}, 0, 0, \ldots\right)$ belong to $\mathbf{l}^1$ for every $n = 1, 2, \ldots$ and the sequence $\{x^{(n)}\}_{n=1}^{\infty}$ is a Cauchy sequence in this space. If the sequence (of sequences) $\{x^{(n)}\}_{n=1}^{\infty}$ converges, then it has to be convergent also coordinate-wise to the element $x^{(0)} = \left(1, \frac{1}{2}, \frac{1}{3}, \ldots, \frac{1}{n}, \frac{1}{n+1}, \ldots\right)$. However, $x^{(0)}$ does not belong to $\mathbf{l}^1$, since $\sum\limits_{n=1}^{\infty} \frac{1}{n} = +\infty$ (see 7.2.1.1, **2.**, p. 459, harmonic series).

12.2.2.2 Complete Metric Spaces

A metric space X is called *complete* if every Cauchy sequence converges in X. Hence, complete metric spaces are the spaces for which the *Cauchy principle*, known from real calculus, is valid: A sequence is convergent if and only if it is a Cauchy sequence. Every closed subspace of a complete metric space (considered as a metric space on its own) is complete. The converse statement is valid in a certain way: If a subspace Y of a (not necessary complete) metric space X is complete, then the set Y is closed in X.

■ Complete metric spaces are, e.g., the spaces: $\mathbf{m}$, $\mathbf{l}^p$ $(1 \le p < \infty)$, $\mathbf{c}$, $\mathcal{B}(T)$, $\mathcal{C}([a,b])$, $\mathcal{C}^{(k)}([a,b])$, $L^p(a,b)$ $(1 \le p < \infty)$.

12.2.2.3 Some Fundamental Theorems in Complete Metric Spaces

The importance of complete metric spaces can be illustrated by a series of theorems and principles, which are known and used in real calculus, and which are to be applied even in the case of infinite dimensional spaces.

1. Theorem on Nested Balls

Let X be a complete metric space. If

$$\overline{B}(x_1;r_1) \supset \overline{B}(x_2;r_2) \supset \cdots \supset \overline{B}(x_n;r_n) \supset \cdots \tag{12.57}$$

is a sequence of nested closed balls with $r_n \longrightarrow 0$, then the intersection of all of those balls is non-empty and consists of only a single point. If this property is valid in some metric space for any sequence satisfying the assumptions, then the metric space is complete.

2. Baire Category Theorem

Let X be a complete metric space and $\{F_k\}_{k=1}^{\infty}$ a sequence of closed sets in X with $\bigcup\limits_{k=1}^{\infty} F_k = \mathrm{X}$. Then there exists at least one index k_0 such that the set F_{k_0} has an interior point.

3. Banach Fixed-Point Theorem

Let F be a non-empty closed subset of a complete metric space (X, ρ). Let T: $X \longrightarrow \mathrm{X}$ be a contracting operator on F, i.e., there exists a constant $q \in [0,1)$ such that

$$\rho(Tx,Ty) \le q\,\rho(x,y) \quad \text{for all} \quad x,y \in F. \tag{12.58}$$

Suppose, if $x \in F$, then $Tx \in F$. Then the following statements are valid:

a) For an arbitrary initial point $x_0 \in F$ the iteration

$$x_{n+1} := Tx_n \quad (n = 0,1,2,\ldots) \tag{12.59}$$

is well defined, i.e., $x_n \in F$ for every n.

b) The iteration sequence $\{x_n\}_{n=0}^{\infty}$ converges to an element $x^* \in F$.

c) $Tx^* = x^*$, i.e., x^* is a fixed point of the operator T. (12.60)

d) The only fixed point of T in F is x^*.

e) The following error estimation is valid:

$$\rho(x^*,x_n) \le \frac{q^n}{1-q}\rho(x_1,x_0). \tag{12.61}$$

The *Banach fixed-point theorem* is sometimes called the *contraction mapping principle*.

12.2.2.4 Some Applications of the Contraction Mapping Principle

1. Iteration Method for Solving a System of Linear Equations

The given linear (n,n) system of equations

$$\begin{array}{l} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_1 = b_1, \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n = b_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n = b_n \end{array} \tag{12.62a}$$

can be transformed according to 19.2.1, p. 955, into the equivalent system

$$\begin{array}{llllll} x_1 & -(1-a_{11})x_1 & +a_{12}x_2 & +\cdots & +a_{1n}x_n & = b_1, \\ x_2 & +a_{21}x_1 & -(1-a_{22})x_2 & +\cdots & +a_{2n}x_n & = b_2, \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ x_n & +a_{n1}x_1 & +a_{n2}x_2 & +\cdots & -(1-a_{nn})x_n & = b_n. \end{array} \tag{12.62b}$$

If the operator T: $\mathbb{F}^n \to \mathbb{F}^n$ is defined by

$$Tx = \left(x_1 - \sum_{k=1}^{n} a_{1k}x_k + b_1, \ldots, x_n - \sum_{k=1}^{n} a_{nk}x_k + b_n \right)^{\mathrm{T}}, \tag{12.63}$$

then the last system is transformed into the fixed-point problem

$$x = Tx \tag{12.64}$$

in the metric space $\mathbb{F}^n$, where an appropriate metric is considered: The Euclidean (12.43), the maximum (12.44) or the absolute value metric $\rho(x, y) = \sum\limits_{k=1}^{n} |x_k - y_k|$ (compare with (12.45)). If one of the numbers

$$\sqrt{\sum_{j,k=1}^{n} |a_{jk}|^2}, \quad \max_{1 \le j \le n} \sum_{k=1}^{n} |a_{jk}|, \quad \max_{1 \le k \le n} \sum_{j=1}^{n} |a_{jk}| \tag{12.65}$$

is smaller than one, then T turns out to be a contracting operator. It has exactly one fixed point according to the Banach fixed-point theorem, which is the componentwise limit of the iteration sequence started from an arbitrary point of $\mathbb{F}^n$.

2. Fredholm Integral Equations

The Fredholm integral equation of second kind (see also 11.2, p. 622)

$$\varphi(x) - \int_a^b K(x, y)\varphi(y)\, dy = f(x), \quad x \in [a, b] \tag{12.66}$$

with a continuous kernel $K(x, y)$ and continuous right-hand side $f(x)$ can be solved by iteration. By means of the operator T: $\mathcal{C}([a, b]) \longrightarrow \mathcal{C}([a, b])$ defined as

$$T\varphi(x) = \int_a^b K(x, y)\varphi(y)\, dy + f(x) \quad \forall\, \varphi \in \mathcal{C}([a, b]), \tag{12.67}$$

it is transformed into a fixed-point problem $T\varphi = \varphi$ in the metric space $\mathcal{C}([a, b])$ (see example **A** in 12.1.2, **4.**, p. 655). If $\max\limits_{a \le x \le b} \int_a^b |K(x, y)|\, dy < 1$, then T is a contracting operator and the fixed-point theorem can be applied. The unique solution is now obtained as the uniform limit of the iteration sequence $\{\varphi_n\}_{n=1}^{\infty}$, where $\varphi_n = T\varphi_{n-1}$, starting with an arbitrary function $\varphi_0(x) \in \mathcal{C}([a, b])$. It is clear that $\varphi_n = T^n\varphi_0$ and the iteration sequence is $\{T^n\varphi_0\}_{n=1}^{\infty}$.

3. Volterra Integral Equations

The Volterra integral equation of second kind (see 11.4, p. 643)

$$\varphi(x) - \int_a^x K(x, y)\varphi(y)\, dy = f(x), \quad x \in [a, b] \tag{12.68}$$

with a continuous kernel and a continuous right-hand side can be solved by means of the Volterra integral operator

$$(V\varphi)(x) := \int_a^x K(x, y)\varphi(y)\, dy \quad \forall\, \varphi \in \mathcal{C}([a, b]) \tag{12.69}$$

and $T\varphi = f + V\varphi$ as the fixed-point problem $T\varphi = \varphi$ in the space $\mathcal{C}([a,b])$.

4. Picard-Lindelöf Theorem

Consider the differential equation

$$\dot{x} = f(t,x) \tag{12.70}$$

with a continuous mapping $f\colon I \times G \longrightarrow \mathbb{R}^n$, where I is an open interval of $\mathbb{R}$ and G is an open domain of $\mathbb{R}^n$. Suppose the function f satisfies a Lipschitz condition with respect to x (see 9.1.1.1, **2.** p. 541), i.e., there is a positive constant L such that

$$\varrho(f(t,x_1), f(t,x_2)) \le L\varrho(x_1,x_2) \quad \forall\,(t,x_1),(t,x_2) \in I \times G, \tag{12.71}$$

where ϱ is the Euclidean metric in $\mathbb{R}^n$. (Using the norm (see 12.3.1, p. 669) and the formula (12.81) $\varrho(x,y) = \|x-y\|$ (12.71) can be written as $\|f(t,x_1) - f(t,x_2)\| \le L \cdot \|x_1 - x_2\|$.) Let $(t_0,x_0) \in I \times G$. Then there are numbers $\beta > 0$ and $r > 0$ such that the set $\Omega = \{(t,x) \in \mathbb{R}\times\mathbb{R}^n\colon |t-t_0| \le \beta,\ \varrho(x,x_0) \le r\}$ lies in $I \times G$. Let $M = \max_\Omega \varrho(f(t,x),0)$ and $\alpha = \min\{\beta, \frac{r}{M}\}$. Then there is a number $b > 0$ such that for each $\tilde{x} \in B = \{x \in \mathbb{R}^n\colon \varrho(x,x_0) \le b\}$, the initial value problem

$$\dot{x} = f(t,x), \quad x(t_0) = \tilde{x} \tag{12.72}$$

has exactly one solution $\varphi(t,\tilde{x})$, i.e., $\dot{\varphi}(t,\tilde{x}) = f(t,\varphi(t,\tilde{x}))$ for $\forall t$ satisfying $|t-t_0| \le \alpha$ and $\varphi(t_0,\tilde{x}) = \tilde{x}$. The solution of this initial value problem is equivalent to the solution of the integral equation

$$\varphi(t,\tilde{x}) = \tilde{x} + \int_{t_0}^{t} f(s,\varphi(s,\tilde{x}))\,ds, \quad t \in [t_0 - \alpha, t_0 + \alpha]. \tag{12.73}$$

If X denotes the closed ball $\{\varphi(t,x)\colon d(\varphi(t,x),x_0) \le r\}$ in the complete metric space $\mathcal{C}([t_0 - \alpha, t_0 + \alpha] \times B;\ \mathbb{R}^n)$ with metric

$$d(\varphi,\psi) = \max_{(t,x)\in\{|t-t_0|\le\alpha\}\times B} \varrho(\varphi(t,x),\psi(t,x)), \tag{12.74}$$

then X is a complete metric space with the induced metric. If the operator $T\colon \mathrm{X} \longrightarrow \mathrm{X}$ is defined by

$$T\varphi(t,x) = \tilde{x} + \int_{t_0}^{t} f(s,\varphi(s,\tilde{x}))\,ds \tag{12.75}$$

then T is a contracting operator and the solution of the integral equation (12.73) is the unique fixed point of T which can be calculated by iteration.

12.2.2.5 Completion of a Metric Space

Every (non-complete) metric space X can be completed; more precisely, there exists a metric space $\tilde{\mathrm{X}}$ with the following properties:

a) $\tilde{\mathrm{X}}$ contains a subspace Y isometric to X (see 12.2.3, **2.**, p. 669).

b) Y is everywhere dense in $\tilde{\mathrm{X}}$.

c) $\tilde{\mathrm{X}}$ is a complete metric space.

d) If Z is any metric space with the properties **a)**–**c)**, then Z and $\tilde{\mathrm{X}}$ are isometric.

The complete metric space, defined uniquely in this way up to isometry, is called the *completion* of the space X.

12.2.3 Continuous Operators

1. Continuous Operators

Let $T\colon \mathrm{X} \longrightarrow \mathrm{Y}$ be a mapping of the metric space $\mathrm{X} = (\mathrm{X},\rho)$ into the metric space $\mathrm{Y} = (\mathrm{Y},\varrho)$. T is called *continuous at the point* $x_0 \in \mathrm{X}$ if for every neighborhood $V = V(y_0)$ of the point $y_0 = T(x_0)$

there is a neighborhood $U = U(x_0)$ such that:

$$T(x) \in V \quad \text{for all} \quad x \in U. \tag{12.76}$$

T is called *continuous on the set* $A \subset \mathrm{X}$ if T is continuous at every point of A. Equivalent properties for T to be continuous on X are:

a) For any point $x \in \mathrm{X}$ and any arbitrary sequence $\{x_n\}_{n=1}^{\infty}$, $x_n \in \mathrm{X}$ with $x_n \longrightarrow x$ there always holds $T(x_n) \longrightarrow T(x)$. Hence $\rho(x_n, x) \to 0$ implies $\varrho(T(x_n), T(x)) \to 0$.

b) For any open subset $G \subset \mathrm{Y}$ the inverse image $T^{-1}(G)$ is an open subset in X.

c) For any closed subset $F \subset \mathrm{Y}$ the inverse image $T^{-1}(F)$ is a closed subset in X.

d) For any subset $A \subset \mathrm{X}$ one has $T(\overline{A}) \subset \overline{T(A)}$.

2. Isometric Spaces

If there is a bijective mapping $T\colon \mathrm{X} \longrightarrow \mathrm{Y}$ for two metric spaces $\mathrm{X} = (\mathrm{X}, \rho)$ and $\mathrm{Y} = (\mathrm{Y}, \varrho)$ such that

$$\rho(x, y) = \varrho(T(x), T(y)) \quad \forall\, x, y \in \mathrm{X}, \tag{12.77}$$

then the spaces X and Y are called *isometric*, and T is called an *isometry*.

12.3 Normed Spaces

12.3.1 Notion of a Normed Space

12.3.1.1 Axioms of a Normed Space

Let X be a vector space over the field $\mathbb{F}$. A function $\|\cdot\|\colon \mathrm{X} \longrightarrow \mathbb{R}^1_+$ is called a *norm* on the vector space X and the pair $\mathrm{X} = (\mathrm{X}, \|\cdot\|)$ is called a *normed space* over the field $\mathbb{F}$, if for arbitrary elements $x, y \in \mathrm{X}$ and for any scalar $\alpha \in \mathbb{F}$ the following properties, the so-called *axioms of a normed space*, are fulfilled:

(**N1**) $\|x\| \geq 0, \quad \text{and} \quad \|x\| = 0 \quad \text{if and only if } x = 0,$ (12.78)

(**N2**) $\|\alpha x\| = |\alpha| \cdot \|x\|$ (homogenity), (12.79)

(**N3**) $\|x + y\| \leq \|x\| + \|y\|$ (triangle inequality). (12.80)

A metric can be introduced by means of

$$\rho(x, y) = \|x - y\|, \quad x, y \in \mathrm{X}, \tag{12.81}$$

in any normed space. The metric (12.81) has the following additional properties which are compatible with the structure of the vector space:

$$\rho(x + z, y + z) = \rho(x, y), \qquad z \in X \tag{12.82a}$$

$$\rho(\alpha x, \alpha y) = |\alpha|\rho(x, y), \quad \alpha \in \mathbb{F}. \tag{12.82b}$$

So, in a normed space there are available both the properties of a vector space and the properties of a metric space. These properties are compatible in the sense of (12.82a) and (12.82b). The advantage is that most of the local investigations can be restricted to the *unit ball*

$$B(0; 1) = \{x \in \mathrm{X} : \|x\| < 1\} \quad \text{or} \quad \overline{B}(0; 1) = \{x \in \mathrm{X} : \|x\| \leq 1\} \tag{12.83}$$

since

$$B(x; r) = \{y \in \mathrm{X} : \|y - x\| < r\} = x + rB(0; 1), \quad \forall\, x \in X \quad \text{and} \quad \forall\, r > 0. \tag{12.84}$$

Moreover, the algebraic operations in a vector space are continuous, i.e.,

$$x_n \to x, \quad y_n \to y, \quad \alpha_n \to \alpha \quad \text{imply}$$

$$x_n + y_n \to x + y, \quad \alpha_n x_n \to \alpha x, \quad \|x_n\| \to \|x\|. \tag{12.85}$$

In normed spaces instead of (12.53) one may write for convergent sequences

$$\|x_n - x_0\| \longrightarrow 0 \quad (n \to \infty). \tag{12.86}$$

12.3.1.2 Some Properties of Normed Spaces

Among the linear metric spaces, those spaces are *normable* (i.e., a norm can be introduced by means of the metric, if one defines $\|x\| = \rho(x, 0)$) whose metric satisfies the conditions (12.82a) and (12.82b).

Two normed spaces X and Y are called *norm isomorphic* if there is a bijective linear mapping $T\colon \mathrm{X} \longrightarrow \mathrm{Y}$ with $\|Tx\| = \|x\|$ for all $x \in \mathrm{X}$. Let $\|\cdot\|_1$ and $\|\cdot\|_2$ be two norms on the vector space X, and denote the corresponding normed spaces by X_1 and X_2, i.e., $\mathrm{X}_1 = (\mathrm{X}, \|\cdot\|_1)$ and $\mathrm{X}_2 = (X, \|\cdot\|_2)$.

The norm $\|\cdot\|_1$ is *stronger* than the norm $\|\cdot\|_2$, if there is a number $\gamma > 0$ such that $\|x\|_2 \le \gamma \|x\|_1$, for all $x \in \mathrm{X}$. In this case, the convergence of a sequence $\{x_n\}_{n=1}^{\infty}$ to x with respect to the stronger norm $\|\cdot\|_1$, i.e., $\|x_n - x\|_1 \longrightarrow 0$, implies the convergence to x with respect to the norm $\|\cdot\|_2$, i.e., $\|x_n - x\|_2 \longrightarrow 0$.

Two norms $\|\cdot\|$ and $\|\cdot\|_1$ are called *equivalent* if there are two numbers $\gamma_1 > 0,\ \gamma_2 > 0$ such that $\forall x \in \mathrm{X}$ there holds $\gamma_1 \|x\| \le \|x\|_1 \le \gamma_2 \|x\|$. In a finite dimensional vector space all norms are equivalent to each other.

A *subspace of a normed space* is a closed linear subspace of the space.

12.3.2 Banach Spaces

A complete normed space is called a *Banach space*. Every normed space X can be completed into a Banach space $\tilde{\mathrm{X}}$ by the completion procedure given in 12.2.2.5, p. 668, and by the natural extension of its algebraic operations and the norm to $\tilde{\mathrm{X}}$.

12.3.2.1 Series in Normed Spaces

In a normed space X *infinite series* can be considered. That means for a given sequence $\{x_n\}_{n=1}^{\infty}$ of elements $x_n \in \mathrm{X}$ a new sequence $\{s_k\}_{k=1}^{\infty}$ is constructed by

$$s_1 = x_1, s_2 = x_1 + x_2, \ldots, s_k = x_1 + \cdots + x_k = s_{k-1} + x_k, \ \ldots . \tag{12.87}$$

If the sequence $\{s_k\}_{k=1}^{\infty}$ is convergent, i.e., $\|s_k - s\| \longrightarrow 0 \ (k \to \infty)$ for some $s \in X$, then a convergent series is defined. The elements $s_1, s_2, \ldots, s_k, \ldots$ are called the partial sums of the series. The limit

$$s = \lim_{k\to\infty} \sum_{n=1}^{k} x_n \tag{12.88}$$

is the *sum* of the series, and it is denoted by $s = \sum\limits_{n=1}^{\infty} x_n$. A series $\sum\limits_{n=1}^{\infty} x_n$ is called *absolutely convergent* if the number series $\sum\limits_{n=1}^{\infty} \|x_n\|$ is convergent. In a Banach space every absolutely convergent series is convergent, and $\|s\| \le \sum\limits_{n=1}^{\infty} \|x_n\|$ holds for its sum s.

12.3.2.2 Examples of Banach Spaces

■ **A :** $\mathbb{F}^n$ with $\|x\| = \left(\sum\limits_{k=1}^{n} |\xi_k|^p \right)^{\frac{1}{p}}$, if $1 \le p < \infty$; $\quad \|x\| = \max\limits_{1\le k\le n} |\xi_k|$, if $p = \infty$. (12.89a)

These normed spaces over the same vector space $\mathbb{F}^n$ are often denoted by $\mathbf{l}^p(n)$ $(1 \le p \le \infty)$. For $1 \le p < \infty$, they are called Euclidean spaces in the case of $\mathbb{F} = \mathbb{R}$, and *unitary spaces* in the case of $\mathbb{F} = \mathbb{C}$.

■ **B :** $\mathbf{m}$ with $\|x\| = \sup\limits_{k} |\xi_k|$. (12.89b)

■ **C :** $\mathbf{c}$ and $\mathbf{c_0}$ with the norm from $\mathbf{m}$. (12.89c)

■ **D :** $\mathbf{l}^p$ with $\|x\| = \|x\|_p = \left(\sum\limits_{n=1}^{\infty} |\xi_n|^p \right)^{\frac{1}{p}}$ $\quad (1 \le p < \infty)$. (12.89d)

■ **E** : $\mathcal{C}([a,b])$ with $\|x\| = \max\limits_{t\in[a,b]} |x(t)|$. (12.89e)

■ **F** : $L^p((a,b))$ $(1 \le p < \infty)$ with $\|x\| = \|x\|_p = \left(\int\limits_a^b |x(t)|^p\, dt \right)^{\frac{1}{p}}$. (12.89f)

■ **G** : $\mathcal{C}^{(k)}([a,b])$ with $\|x\| = \sum\limits_{l=0}^{k} \max\limits_{t\in[a,b]} |x^{(l)}(t)|$, where $x^{(0)}(t)$ stands for $x(t)$. (12.89g)

12.3.2.3 Sobolev Spaces

Let $\Omega \subset \mathbb{R}^n$ be a bounded domain, i.e., an open connected set, with a sufficiently smooth boundary $\partial\Omega$. For $n = 1$ or $n = 2, 3$ one can imagine Ω being something similar to an interval (a, b) or a bounded convex set.

A function $f\colon \overline{\Omega} \longrightarrow \mathbb{R}$ is k-times continuously differentiable on the closed domain $\overline{\Omega}$ if f is k-times continuously differentiable on Ω and each of its partial derivatives has a finite limit on the boundary, i.e., if x approaches an arbitrary point of $\partial\Omega$. In other words, all partial derivatives can be continuously extended on the boundary of Ω, i.e., each partial derivative is a continuous function on $\overline{\Omega}$. In this vector space (for $p \in [1, \infty)$) and with the Lebesgue measure λ in $\mathbb{R}^n$ (see example **C** in 12.9.1, **2.**, p. 695) the following norm is defined:

$$\|f\|_{k,p} = \|f\| = \left(\int\limits_{\Omega} |f(x)|^p\, d\lambda + \sum_{1\le|\alpha|\le k} \int\limits_{\Omega} |D^\alpha f|^p\, d\lambda \right)^{\frac{1}{p}}. \qquad (12.90)$$

The resulting normed space is denoted by $\tilde{W}^{k,p}(\Omega)$ or also by $\tilde{W}_p^k(\Omega)$ (in contrast to the space $\mathcal{C}^{(k)}([a,b])$ which has a quite different norm). Here α means a *multi-index*, i.e., an ordered n-tuple $(\alpha_1, \ldots, \alpha_n)$ of non-negative integers, where the sum of the components of α is denoted by $|\alpha| = \alpha_1 + \alpha_2 + \cdots + \alpha_n$. For a function $f(x) = f(\xi_1, \ldots, \xi_n)$ with $x = (\xi_1, \ldots, \xi_n) \in \overline{\Omega}$ the brief notation is used as in (12.90):

$$D^\alpha f = \frac{\partial^{|\alpha|} f}{\partial\xi_1^{\alpha_1} \cdots \partial\xi_n^{\alpha_n}}. \qquad (12.91)$$

The normed space $\tilde{W}^{k,p}(\Omega)$ is not complete. Its completion is denoted by $W^{k,p}(\Omega)$ or in the case of $p = 2$ by $\mathbb{H}^k(\Omega)$ and it is called a *Sobolev space*.

12.3.3 Ordered Normed Spaces

1. Cones in a Normed Space

Let X be a real normed space with the norm $\|\cdot\|$. A cone $X_+ \subset X$ (see 12.1.4.2, p. 657) is called *solid*, if X_+ contains a ball (with positive radius), or equivalently, X_+ contains at least one interior point.

■ The usual cones are solid in the spaces $\mathbb{R}$, $\mathcal{C}([a,b])$, $\mathbf{c}$, but in the spaces $L^p((a,b))$ and $\mathbf{l}^p$ $(1 \le p < \infty)$ they are not solid.

A cone X_+ is called *normal* if the norm in X is *semi-monotonic*, i.e., there exists a constant $M > 0$ such that

$$0 \le x \le y \implies \|x\| \le M\|y\|. \qquad (12.92)$$

If X is a Banach space ordered by a cone X_+, then every (o)-interval is bounded with respect to the norm if and only if the cone X_+ is normal.

■ The cones of the vectors with non-negative components and of the non-negative functions in the spaces $\mathbb{R}^n$, $\mathbf{m}$, $\mathbf{c}$, $\mathbf{c_0}$, $\mathcal{C}$, $\mathbf{l}^p$ and L^p, respectively, are normal.

A cone is called *regular* if every monotonically increasing sequence which is bounded above,

$$x_1 \le x_2 \le \cdots \le x_n \le \cdots \le z \qquad (12.93)$$

is a Cauchy sequence in X. In a Banach space every closed regular cone is normal.

■ The cones in $\mathbb{R}^n$, $\mathbf{l}^p$ and L^p for $1 \le p < \infty$ are regular, but in $\mathcal{C}$ and $\mathbf{m}$ they are not.

2. Normed Vector Lattices and Banach Lattices

Let X be a vector lattice, which is a normed space at the same time. X is called a *normed lattice* or *normed vector lattice* (see [12.15], [12.19], [12.22], [12.23]), if the norm satisfies the condition

$$|x| \le |y| \quad \text{implies} \quad \|x\| \le \|y\| \quad \forall\, x, y \in \mathrm{X} \quad \text{(monotonicity of the norm).} \tag{12.94}$$

A complete (with respect to the norm) normed lattice is called a *Banach lattice.*

■ The spaces $\mathcal{C}([a,b])$, L^p, $\mathbf{l}^p$, $\mathcal{B}([a,b])$ are Banach lattices.

12.3.4 Normed Algebras

A vector space X over $\mathbb{F}$ is called an *algebra*, if in addition to the operations defined in the vector space X and satisfying the axioms **(V1)**–**(V8)** (see 12.1.1, p. 654), a product $x \cdot y \in \mathrm{X}$ is defined for every two elements $x, y \in \mathrm{X}$ (or with a simplified notation by a product xy), such that for arbitrary $x, y, z \in \mathrm{X}$ and $\alpha \in \mathbb{F}$ the following conditions are satisfied:

(**A1**) $x(yz) = (xy)z$, (12.95)

(**A2**) $x(y + z) = xy + xz$, (12.96)

(**A3**) $(x + y)z = xz + yz$, (12.97)

(**A4**) $\alpha(xy) = (\alpha x)y = x(\alpha y)$. (12.98)

An algebra is *commutative* if $xy = yx$ holds for two arbitrary elements x, y. A linear operator (see (12.21), p. 658) $T\colon \mathrm{X} \longrightarrow \mathrm{Y}$ of the algebra X into the algebra Y is called an *algebra homomorphism* if for any $x_1, x_2 \in \mathrm{X}$:

$$T(x_1 \cdot x_2) = Tx_1 \cdot Tx_2. \tag{12.99}$$

An algebra X is called a *normed algebra* or a *Banach algebra* if it is a normed vector space or a Banach space and the norm has the additional property

$$\|x \cdot y\| \le \|x\| \cdot \|y\|. \tag{12.100}$$

In a normed algebra all the operations are continuous, i.e., additionally to (12.85), if $x_n \longrightarrow x$ and $y_n \longrightarrow y$, then also $x_n y_n \longrightarrow xy$ (see [12.20]).

Every normed algebra can be completed to a Banach algebra, where the product is extended to the norm completion with respect to (12.100).

■ **A:** $\mathcal{C}([a,b])$ with the norm (12.89e) and the usual (pointwise) product of continuous functions.

■ **B:** The vector space $W([0, 2\pi])$ of all complex-valued functions $x(t)$ continuous on $[0, 2\pi]$ and having an absolutely convergent Fourier series expansion, i.e.,

$$x(t) = \sum_{n=-\infty}^{\infty} c_n e^{int}, \tag{12.101}$$

with the norm $\|x\| = \sum_{n=-\infty}^{\infty} |c_n|$ and the usual multiplication.

■ **C:** The space $L(\mathrm{X})$ of all bounded linear operators on the normed space X with the operator norm and the usual algebraic operations (see 12.5.1.2, p. 677), where the product TS of two operators is defined as the sequential application, i.e., $TS(x) = T(S(x))$, $x \in \mathrm{X}$.

■ **D:** The space $L^1(-\infty, \infty)$ of all measurable and absolutely integrable functions on the real axis (see 12.9, p. 693) with the norm

$$\|x\| = \int_{-\infty}^{\infty} |x(t)|\, dt \tag{12.102}$$

is a Banach algebra if the multiplication is defined as the convolution $(x * y)(t) = \int_{-\infty}^{\infty} x(t-s)y(s)\, ds$.

12.4 Hilbert Spaces

12.4.1 Notion of a Hilbert Space

12.4.1.1 Scalar Product

A vector space V over a field $\mathbb{F}$ (mostly $\mathbb{F} = \mathbb{C}$) is called a *space with scalar product* or an *inner product space* or *pre-Hilbert space* if to every pair of elements $x, y \in \mathrm{V}$ there is assigned a number $(x, y) \in \mathbb{F}$ (the scalar product of x and y), such that the *axioms of the scalar product* are satisfied, i.e., for arbitrary $x, y, z \in \mathrm{V}$ and $\alpha \in \mathbb{F}$:

(**H1**) $(x, x) \geq 0$, (i.e., (x, x) is real), and $(x, x) = 0$ if and only if $x = 0$, (12.103)

(**H2**) $(\alpha x, y) = \alpha(x, y)$, (12.104)

(**H3**) $(x + y, z) = (x, z) + (y, z)$, (12.105)

(**H4**) $(x, y) = \overline{(y, x)}$. (12.106)

(Here $\overline{\omega}$ denotes the conjugate of the complex number ω, which is denoted by ω^* in (1.133c). Sometimes the notation of a scalar product is $\langle x, y \rangle$.)

In the case of $\mathbb{F} = \mathbb{R}$, i.e., in a real vector space, **(H4)** means the commutativity of the scalar product. Some further properties follow from the axioms:

$$(x, \alpha y) = \bar{\alpha}(x, y) \quad \text{and} \quad (x, y + z) = (x, y) + (x, z). \tag{12.107}$$

12.4.1.2 Unitary Spaces and Some of their Properties

In a pre-Hilbert space $\mathbb{H}$ a norm can be introduced by means of the scalar product as follows:

$$\|x\| = \sqrt{(x, x)} \quad (x \in \mathbb{H}). \tag{12.108}$$

A normed space $\mathbb{H} = (\mathbb{H}, \|\cdot\|)$ is called *unitary* if there is a scalar product satisfying (12.108). Based on the previous properties of scalar products and (12.108) in unitary spaces the following facts are valid:

a) Triangle Inequality:

$$\|x + y\| \leq \|x\| + \|y\|. \tag{12.109}$$

b) Cauchy-Schwarz Inequality or Schwarz-Buniakowski Inequality (see also 1.4.2.9, p. 31):

$$|(x, y)| \leq \sqrt{(x, x)}\sqrt{(y, y)}\,. \tag{12.110}$$

c) Parallelogram Identity: This characterizes the unitary spaces among the normed spaces:

$$\|x + y\|^2 + \|x - y\|^2 = 2\left(\|x\|^2 + \|y\|^2\right). \tag{12.111}$$

d) Continuity of the Scalar Product:

$$x_n \to x,\ y_n \to y \quad \text{imply} \quad (x_n, y_n) \to (x, y). \tag{12.112}$$

12.4.1.3 Hilbert Space

A complete unitary space is called a *Hilbert space*. Since Hilbert spaces are also Banach spaces, they possess their properties (see 12.3.1, p. 669; 12.3.1.2, p. 670; 12.3.2, p. 670). In addition they have the properties of unitary spaces 12.4.1.2, p. 673. A subspace of a Hilbert space is a closed linear subspace.

■ **A:** $\mathbf{l}^2(n)$, $\mathbf{l}^2$ and $L^2((a, b))$ with the scalar products

$$(x, y) = \sum_{k=1}^{n} \xi_k \overline{\eta_k},\ (x, y) = \sum_{k=1}^{\infty} \xi_k \overline{\eta_k} \quad \text{and} \quad (x, y) = \int_a^b x(t)\overline{y(t)}\, dt. \tag{12.113}$$

■ **B:** The space $\mathbb{H}^2(\Omega)$ with the scalar product

$$(f, g) = \int_\Omega f(x)\overline{g(x)}\, dx + \sum_{1 \leq |\alpha| \leq k} \int_\Omega D^\alpha f(x)\overline{D^\alpha g(x)}\, dx. \tag{12.114}$$

■ **C:** Let $\varphi(t)$ be a measurable positive function on $[a, b]$. The complex space $L^2((a, b), \varphi)$ of all measurable functions, which are quadratically integrable with the *weight* function φ on (a, b), is a Hilbert space if the scalar product is defined as

$$(x, y) = \int_a^b x(t)\overline{y(t)}\varphi(t)\,dt. \tag{12.115}$$

12.4.2 Orthogonality

Two elements x, y of a Hilbert space $\mathbb{H}$ are called *orthogonal* (denoted by $x \perp y$) if $(x, y) = 0$ (the notions of this paragraph also make sense in pre-Hilbert spaces and in unitary spaces). For an arbitrary subset $A \subset \mathbb{H}$, the set

$$A^\perp = \{x \in \mathbb{H}\colon (x, y) = 0 \quad \forall\, y \in A\} \tag{12.116}$$

of all vectors which are orthogonal to each vector in A is a (closed linear) subspace of $\mathbb{H}$ and it is called the *orthogonal space* to A or the *orthogonal complement* of A. The notation $A \perp B$ means that $(x, y) = 0$ for all $x \in A$ and $y \in B$. If A consists of a single element x, then the notation $x \perp B$ is used.

12.4.2.1 Properties of Orthogonality

The zero vector is orthogonal to every vector of $\mathbb{H}$. The following statements hold:

a) $x \perp y$ and $x \perp z$ imply $x \perp (\alpha y + \beta z)$ for any $\alpha, \beta \in \mathbb{C}$.

b) From $x \perp y_n$ and $y_n \to y$ it follows that $x \perp y$.

c) $x \perp A$ if and only if $x \perp \overline{lin(A)}$, where $\overline{lin(A)}$ denotes the *closed linear hull* of the set A.

d) If $x \perp A$ and A is a *fundamental set*, i.e., $lin(A)$ is everywhere dense in $\mathbb{H}$, then $x = 0$.

e) Pythagoras Theorem: If the elements $x_1, \ldots, x_n$ are pairwise orthogonal, that is $x_k \perp x_l$ for all $k \neq l$, then

$$\|\sum_{k=1}^{n} x_k\|^2 = \sum_{k=1}^{n} \|x_k\|^2. \tag{12.117}$$

f) Projection Theorem: If $\mathbb{H}_0$ is a subspace of $\mathbb{H}$, then each vector $x \in \mathbb{H}$ can be written uniquely as

$$x = x' + x'', \quad x' \in \mathbb{H}_0,\ x'' \perp \mathbb{H}_0. \tag{12.118}$$

g) Approximation Problem: Furthermore, the equation $\|x'\| = \rho(x, \mathbb{H}_0) = \inf_{y\in\mathbb{H}_0}\{\|x - y\|\}$ holds, and so the problem

$$\|x - y\| \to \inf, \quad y \in \mathbb{H}_0 \tag{12.119}$$

has the unique solution x' in $\mathbb{H}_0$. In this statement $\mathbb{H}_0$ can be replaced by a convex closed non-empty subset of $\mathbb{H}$.
The element x' is called the *projection* of the element x on $\mathbb{H}_0$. It has the smallest distance from x (to $\mathbb{H}_0$), and the space $\mathbb{H}$ can be decomposed: $\mathbb{H} = \mathbb{H}_0 \oplus \mathbb{H}_0^\perp$.

12.4.2.2 Orthogonal Systems

A set $\{x_\xi\colon \xi \in \Xi\}$ of vectors from $\mathbb{H}$ is called an *orthogonal* system if it does not contain the zero vector and $x_\xi \perp x_\eta$, $\xi \neq \eta$, hence $(x_\xi, x_\eta) = \delta_{\xi\eta}$ holds, where

$$\delta_{\xi\eta} = \begin{cases} 1 & \text{for } \xi = \eta, \\ 0 & \text{for } \xi \neq \eta \end{cases} \tag{12.120}$$

denotes the Kronecker symbol (see 4.1.2, **10.**, p. 271). An orthogonal system is called *orthonormal* if in addition $\|x_\xi\| = 1\ \forall\, \xi$.
In a separable Hilbert space an orthogonal system may contain at most countably many elements. Therefore $\Xi = \mathbb{N}$ is assumed from now on.

■ **A:** The system

$$\frac{1}{\sqrt{2\pi}},\ \frac{1}{\sqrt{\pi}}\cos t,\ \frac{1}{\sqrt{\pi}}\sin t,\ \frac{1}{\sqrt{\pi}}\cos 2t,\ \frac{1}{\sqrt{\pi}}\sin 2t,\ \ldots \tag{12.121}$$

in the real space $L^2((-\pi,\pi))$ and the system

$$\frac{1}{\sqrt{2\pi}}e^{int} \quad (n = 0, \pm 1, \pm 2, \ldots) \tag{12.122}$$

in the complex space $L^2((-\pi,\pi))$ are orthonormal systems. Both of these systems are called *trigonometric*.

■ **B:** The Legendre polynomials of the first kind (see 9.1.2.6, **2.**, p. 566)

$$P_n(t) = \frac{d^n}{d\,t^n}[(t^2-1)^n] \quad (n = 0, 1, \ldots) \tag{12.123}$$

form an orthogonal system of elements in the space $L^2((-1,1))$. The corresponding orthonormal system is

$$\tilde{P}_n(t) = \sqrt{n + \frac{1}{2}}\,\frac{1}{(2n)!!}\,P_n(t). \tag{12.124}$$

■ **C:** The Hermite polynomials (see 9.1.2.6, **6.**, p. 568 and 9.2.4, **3.**, 602) according to the second definition of the Hermite differential equation (9.66b)

$$H_n(t) = e^{t^2}\frac{d^n}{dt^n}e^{-t^2} \quad (n = 0, 1, \ldots) \tag{12.125}$$

form an orthogonal system in the space $L^2((-\infty,\infty))$.

■ **D:** The Laguerre polynomials form an orthogonal system (see 9.1.2.6, **5.**, p. 568) in the space $L^2((0,\infty))$.

Every orthogonal system is linearly independent, since the zero vector was excluded. Conversely, if $x_1, x_2, \ldots x_n, \ldots$ is a system of linearly independent elements in a Hilbert space $\mathbb{H}$, then there exist vectors $e_1, e_2, \ldots, e_n, \ldots$, obtained by the *Gram–Schmidt orthogonalization method* (see 4.6.2.2, **1.**, p. 316) which form an orthonormal system. They span the same subspace, and by the method they are determined up to a scalar factor with modulus 1.

12.4.3 Fourier Series in Hilbert Spaces

12.4.3.1 Best Approximation

Let $\mathbb{H}$ be a separable Hilbert space and

$$\{e_n\colon\ n = 1, 2, \ldots\} \tag{12.126}$$

a fixed orthonormal system in $\mathbb{H}$. For an element $x \in \mathbb{H}$ the numbers $c_n = (x, e_n)$ are called the *Fourier coefficients* of x with respect to the system (12.126). The (formal) series

$$\sum_{n=1}^{\infty} c_n e_n \tag{12.127}$$

is called the *Fourier series* of the element x with respect to the system (12.126) (see 7.4.1.1, **1.**, p. 474). The n-th partial sum of the Fourier series of an element x has the property of the *best approximation*, i.e., for fixed n, the n-th partial sum of the Fourier series

$$\sigma_n = \sum_{k=1}^{n}(x, e_k)e_k \tag{12.128}$$

gives the smallest value of $\|x - \sum_{k=1}^{n} \alpha_k e_k\|$ among all vectors of $\mathbb{H}_n = lin(\{e_1, \ldots, e_n\})$. Furthermore, $x - \sigma_n$ is orthogonal to $\mathbb{H}_n$, and there holds the *Bessel inequality*:

$$\sum_{n=1}^{\infty} |c_n|^2 \leq \|x\|^2, \quad c_n = (x, e_n) \quad (n = 1, 2, \ldots). \tag{12.129}$$

12.4.3.2 Parseval Equation, Riesz-Fischer Theorem

The Fourier series of an arbitrary element $x \in \mathbb{H}$ is always convergent. Its sum is the projection of the element x onto the subspace $\mathbb{H}_0 = \overline{lin(\{e_n\}_{n=1}^{\infty})}$. If an element $x \in \mathbb{H}$ has the representation $x = \sum_{n=1}^{\infty} \alpha_n e_n$, then α_n are the Fourier coefficients of x $(n = 1, 2, \ldots)$. If $\{\alpha_n\}_{n=1}^{\infty}$ is an arbitrary sequence of numbers with the property $\sum_{n=1}^{\infty} |\alpha_n|^2 < \infty$, then there is a unique element x in $\mathbb{H}$, whose Fourier coefficients are equal to α_n and for which the Parseval equation holds:

$$\sum_{n=1}^{\infty} |(x, e_n)|^2 = \sum_{n=1}^{\infty} |\alpha_n|^2 = \|x\|^2 \qquad \text{(Riesz–Fischer theorem)}. \tag{12.130}$$

An orthonormal system $\{e_n\}$ in $\mathbb{H}$ is called *complete* if there is no non-zero vector y orthogonal to every e_n; it is called a *basis* if every vector $x \in \mathbb{H}$ has the representation $x = \sum_{n=1}^{\infty} \alpha_n e_n$, i.e., $\alpha_n = (x, e_n)$ and x is equal to the sum of its Fourier series. In this case, one also says that x has a Fourier expansion. The following statements are equivalent:

a) $\{e_n\}$ is a fundamental set in $\mathbb{H}$.

b) $\{e_n\}$ is complete in $\mathbb{H}$.

c) $\{e_n\}$ is a basis in $\mathbb{H}$.

d) For $\forall\, x, y \in \mathbb{H}$ with the corresponding Fourier coefficients c_n and d_n $(n = 1, 2, \ldots)$ there holds

$$(x, y) = \sum_{n=1}^{\infty} c_n \overline{d_n}. \tag{12.131}$$

e) For every vector $x \in \mathbb{H}$, the Parseval equation (12.130) holds.

■ **A:** The trigonometric system (12.121) is a basis in the space $L^2((-\pi, \pi))$.

■ **B:** The system of the normalized Legendre polynomials (12.124) $\tilde{P}_n(t)$ $(n = 0, 1, \ldots)$ is complete and consequently a basis in the space $L^2((-1, 1))$.

12.4.4 Existence of a Basis, Isomorphic Hilbert Spaces

In every separable Hilbert space there exits a basis. From this fact it follows that every orthonormal system can be completed to a basis.

Two Hilbert spaces $\mathbb{H}_1$ and $\mathbb{H}_2$ are called *isometric* or *isomorphic* as Hilbert spaces if there is a linear bijective mapping T: $\mathbb{H}_1 \longrightarrow \mathbb{H}_2$ with the property $(Tx, Ty)_{\mathbb{H}_2} = (x, y)_{\mathbb{H}_1}$ (that is, it preserves the scalar product and because of (12.108) also the norm). Any two arbitrary infinite dimensional separable Hilbert spaces are isometric, in particular every such space is isometric to the separable space $\mathbf{l}^2$.

12.5 Continuous Linear Operators and Functionals

12.5.1 Boundedness, Norm and Continuity of Linear Operators

12.5.1.1 Boundedness and the Norm of Linear Operators

Let $X = (X, \|\cdot\|)$ and $Y = (Y, \|\cdot\|)$ be normed spaces. In the following discussion the index X in the notation $\|\cdot\|_X$, which emphasizes being in the space X, is omitted, because from the text it will be always clear, which norms and spaces are considered. An arbitrary operator $T\colon X \longrightarrow Y$ is called *bounded* if there is a real number $\lambda > 0$ such that

$$\|T(x)\| \le \lambda\|x\| \quad \forall\, x \in X. \tag{12.132}$$

A bounded operator with a constant λ "stretches" every vector at most λ times and it transforms every bounded set of X into a bounded set of Y, in particular the image of the unit ball of X is bounded in Y. This last property is characteristic of bounded linear operators. A linear operator is continuous (see 12.2.3, p. 668) if and only if it is bounded.

The smallest constant λ, for which (12.132) still holds, is called the *norm of the operator* T and it is denoted by $\|T\|$, i.e.,

$$\|T\| := \inf\{\lambda > 0 : \|Tx\| \le \lambda\|x\|,\ x \in X\}. \tag{12.133}$$

For a continuous linear operator the following equalities hold:

$$\|T\| = \sup_{\|x\|\le 1}\|Tx\| = \sup_{\|x\|<1}\|Tx\| = \sup_{\|x\|=1}\|Tx\| \tag{12.134}$$

and, furthermore, the following estimation holds

$$\|Tx\| \le \|T\|\cdot\|x\| \quad \forall\, x \in X. \tag{12.135}$$

■ Let T be the operator in the space $\mathcal{C}([a,b])$ with the norm (12.89e), defined by the integral

$$(Tx)(s) = y(s) = \int_a^b K(s,t)x(t)\,dt \quad (s \in [a,b]), \tag{12.136}$$

where $K(s,t)$ is a (complex-valued) continuous function on the rectangle $\{a \le s, t \le b\}$. Then T is a bounded linear operator, which maps $\mathcal{C}([a,b])$ into $\mathcal{C}([a,b])$. Its norm is

$$\|T\| = \max_{s\in[a,b]}\int_a^b |K(s,t)|\,dt. \tag{12.137}$$

12.5.1.2 The Space of Linear Continuous Operators

The sum $U = S + T$ and the multiple αT of two linear (continuous) operators $S, T\colon X \longrightarrow Y$ are defined point-wise:

$$U(x) = S(x) + T(x), \quad (\alpha T)(x) = \alpha\cdot T(x), \quad \forall\, x \in X \text{ and } \forall\, \alpha \in \mathbb{F}. \tag{12.138}$$

The set $L(X, Y)$, often denoted by $B(X, Y)$, of all linear continuous operators T from X into Y equipped with the operations (12.138) is a vector space, where $\|T\|$ (12.133) turns out to be a norm on it. So, $L(X, Y)$ is a normed space and even a Banach space if Y is a Banach space. So the axioms **(V1)**–**(V8)** and **(N1)**–**(N3)** are satisfied (see 12.1.1, p. 654, 12.3.1, p. 669).

If $Y = X$, then a product can be defined for two arbitrary elements $S, T \in L(X, X) = L(X) = B(X)$ as

$$(ST)(x) = S(Tx) \quad (\forall\, x \in X), \tag{12.139}$$

which satisfies the axioms **(A1)**–**(A4)** from 12.3.4, p. 672, and also the compatibility condition (12.100) with the norm. $L(X)$ is in general a non-commutative normed algebra, and if X is a Banach space, then it is a Banach algebra. Then for every operator $T \in L(X)$ its powers are defined by

$$T^0 = I,\ T^n = T^{n-1}T \quad (n = 1, 2, \ldots), \tag{12.140}$$

where I is the identity operator $Ix = x, \quad \forall\, x \in X$. Then

$$\|T^n\| \le \|T\|^n \quad (n = 0, 1, \ldots), \tag{12.141}$$

and furthermore there always exists the (finite) limit

$$r(T) = \lim_{n\to\infty} \sqrt[n]{\|T^n\|}\,, \tag{12.142}$$

which is called the *spectral radius* of the operator T and satisfies the relations

$$r(T) \le \|T\|, \quad r(T^n) = [r(T)]^n, \quad r(\alpha T) = |\alpha| r(T), \quad r(T) = r(T^*), \tag{12.143}$$

where T^* is the adjoint operator to T (see 12.6, p. 684, and (12.159)). If $L(\mathrm{X})$ is complete, then for $|\lambda| > r(T)$, the operator $(\lambda I - T)^{-1}$ has the representation in the form of a Neumann series

$$(\lambda I - T)^{-1} = \lambda^{-1} I + \lambda^{-2} T + \ldots + \lambda^{-n} T^{n-1} + \ldots\,, \tag{12.144}$$

which is convergent for $|\lambda| > r(T)$ in the operator norm on $L(\mathrm{X})$.

12.5.1.3 Convergence of Operator Sequences

1. Point-wise Convergence

of a sequence of linear continuous operators $T_n\colon \mathrm{X} \longrightarrow \mathrm{Y}$ to an operator $T\colon \mathrm{X} \longrightarrow \mathrm{Y}$ means that:

$$T_n x \longrightarrow Tx \quad \text{in Y for each } x \in \mathrm{X}. \tag{12.145}$$

2. Uniform Convergence

The usual norm-convergence of a sequence of operators $\{T_n\}_{n=1}^{\infty}$ in a space $L(\mathrm{X}, \mathrm{Y})$ to T, i.e.,

$$\|T_n - T\| = \sup_{\|x\|\le 1} \|T_n x - Tx\| \longrightarrow 0 \quad (n \to \infty) \tag{12.146}$$

is the *uniform* convergence on the unit ball of X. It implies point-wise convergence, while the converse statement is not true in general.

3. Applications

The convergence of quadrature formulas when the number n of interpolation nodes tends to ∞, the performance principle of summation, limiting methods, etc.

12.5.2 Linear Continuous Operators in Banach Spaces

Now X and Y are supposed to be Banach spaces.

1. Banach-Steinhaus Theorem (Uniform Boundedness Principle)

The theorem characterizes the point-wise convergence of a sequence $\{T_n\}$ of linear continuous operators T_n to some linear continuous operator by the conditions:

a) For every element from an everywhere dense subset $D \subset \mathrm{X}$, the sequence $\{T_n x\}$ has a limit in Y,

b) there is a constant C such that $\|T_n\| \le C,\ \forall\, n$.

2. Open Mappings Theorem

The theorem tells us that a linear continuous operator mapping from X onto Y is *open*, i.e., the image $T(G)$ of every open set G from X is an open set in Y.

3. Closed Graph Theorem

An operator $T\colon D_T \longrightarrow \mathrm{Y}$ with $D_T \subset \mathrm{X}$ is called *closed* if $x_n \in D_T$, $x_n \to x_0$ in X and $Tx_n \to y_0$ in Y imply $x_0 \in D_T$ and $y_0 = Tx_0$. A necessary and sufficient condition is that the graph of the operator T in the space $\mathrm{X} \times \mathrm{Y}$, i.e., the set

$$\Gamma_T = \{(x, Tx)\colon\ x \in D_T\} \tag{12.147}$$

is closed, where here (x, y) denotes an element of the set $\mathrm{X} \times \mathrm{Y}$.
If T is a closed operator with a closed domain D_T, then T is continuous.

4. Hellinger-Toeplitz Theorem

Let T be a linear operator in a Hilbert space $\mathbb{H}$. If $(x, Ty) = (Tx, y)$ for every $x, y \in \mathbb{H}$, then T is continuous (here (x, Ty) denotes the scalar product in $\mathbb{H}$).

5. Krein-Losanovskij Theorem on the Continuity of Positive Linear Operators

If $\mathrm{X} = (\mathrm{X}, \mathrm{X}_+, \|\cdot\|)$ and $\mathrm{Y} = (\mathrm{Y}, \mathrm{Y}_+, \|\cdot\|)$ are ordered normed spaces, where X_+ is a generating cone, then the set $L_+(\mathrm{X}, \mathrm{Y})$ of all positive linear and continuous operators T, i.e., $T(\mathrm{X}_+) \subset \mathrm{Y}_+$, is a cone in $L(\mathrm{X}, \mathrm{Y})$. The *theorem of Krein and Losanovskij* asserts (see [12.17]): If X and Y are ordered Banach spaces with closed cones X_+ and Y_+, and X_+ is a generating cone, then the positivity of a linear operator implies its continuity.

6. Inverse Operator

Let X and Y be arbitrary normed spaces and let $T\colon\ \mathrm{X} \longrightarrow \mathrm{Y}$ be a linear, not necessarily continuous operator. T has a continuous inverse $T^{-1}\colon\ \mathrm{Y} \longrightarrow \mathrm{X}$, if $T(\mathrm{X}) = \mathrm{Y}$ and there exists a constant $m > 0$ such that $\|Tx\| \ge m\|x\|$ for each $x \in X$. Then $\|T^{-1}\| \le \dfrac{1}{m}$. The situation considered here is less general than that in (12.22) (see 12.1.5.2, p. 659), since there may be $D \ne \mathrm{X}$ and $\mathcal{R}(T) \ne \mathrm{Y}$.
In the case of Banach spaces X, Y the following theorem is valid:

7. Banach Theorem on the Continuity of the Inverse Operator

If T is a linear continuous bijective operator from X onto Y, then the inverse operator T^{-1} is also continuous.

An important application is, e.g., the continuity of $(\lambda I - T)^{-1}$ given the injectivity and surjectivity of $\lambda I - T$. This fact has importance in investigating the spectrum of an operator (see 12.5.3.2, p. 680). It also applies to the

8. Continuous Dependence of the Solution

on the right-hand side and also on the initial data of initial value problems for linear differential equations. This fact is demonstrated by the following example.

■ The initial value problem

$$\ddot{x}(t) + p_1(t)\dot{x}(t) + p_2(t)x(t) = q(t), \quad t \in [a,b], \quad x(t_0) = \xi,\ \dot{x}(t_0) = \dot{\xi}, \quad t_0 \in [a,b] \tag{12.148a}$$

with coefficients $p_1(t), p_2(t) \in \mathcal{C}([a,b])$ has exactly one solution x from $\mathcal{C}^2([a,b])$ for every right-hand side $q(t) \in \mathcal{C}([a,b])$ and for every pair of numbers $\xi,\ \dot{\xi}$. The solution x depends continuously on $q(t),\ \xi$ and $\dot{\xi}$ in the following sense. If $q_n(t) \in \mathcal{C}([a,b]), \xi_n,\ \dot{\xi}_n \in \mathbb{R}^1$ are given and $x_n \in \mathcal{C}([a,b])$ denotes the solution of

$$\ddot{x}_n(t) + p_1(t)\dot{x}_n(t) + p_2(t)x_n(t) = q_n(t), \quad x_n(a) = \xi_n,\ \dot{x}_n(a) = \dot{\xi}_n, \tag{12.148b}$$

for $n = 1, 2, \ldots$, then:

$$\left.\begin{array}{r} q_n(t) \to q(t) \ \text{ in } \ \mathcal{C}([a,b]), \\ \xi_n \to \xi, \\ \dot{\xi}_n \to \dot{\xi}, \end{array}\right\} \quad \text{implies } x_n \to x \text{ in the space } \mathcal{C}^2([a,b]). \tag{12.148c}$$

9. Method of Successive Approximation

to solve an equation of the form

$$x - Tx = y \tag{12.149}$$

with a continuous linear operator T in a Banach space X for a given y. This method starts with an arbitrary initial element x_0, and constructs a sequence $\{x_n\}$ of approximating solutions by the formula

$$x_{n+1} = y + Tx_n \quad (n = 0, 1, \ldots). \tag{12.150}$$

This sequence converges to the solution x^* in X of (12.149). The convergence of the method, i.e., $x_n \to x^*$, is based on the convergence of the series (12.144) with $\lambda = 1$.
Let $\|T\| \le q < 1$. Then the following statements are valid:

a) The operator $I - T$ has a continuous inverse with $\|(I-T)^{-1}\| \le \dfrac{1}{1-q}$, and (12.149) has exactly one solution for each y.

b) The series (12.144) converges and its sum is the operator $(I-T)^{-1}$.

c) The method (12.150) converges to the unique solution x^* of (12.149) for any initial element x_0, if the series (12.144) converges. Then the following estimation holds:

$$\|x_n - x^*\| \le \frac{q^n}{1-q}\|Tx_0 - x_0\| \quad (n = 1, 2, \ldots). \tag{12.151}$$

Equations of the type

$$x - \mu Tx = y, \quad \lambda x - Tx = y, \quad \mu, \lambda \in \mathbb{F} \tag{12.152}$$

can be handled in an analogous way (see 11.2.2, p. 625, and [12.8]).

12.5.3 Elements of the Spectral Theory of Linear Operators

12.5.3.1 Resolvent Set and the Resolvent of an Operator

For an investigation of the solvability of equations one tries to rewrite the problem in the form

$$(I - T)x = y \tag{12.153}$$

with some operator T having a possible small norm. This is especially convenient for using a functional analytic method because of (12.143) and (12.144). In order to handle large values of $\|T\|$ as well, it is necessary to investigate the whole family of equations

$$(\lambda I - T)x = y \quad x \in \mathrm{X}, \text{ with } \lambda \in \mathbb{C} \tag{12.154}$$

in a complex Banach space X. Let T be a linear, but in general not a bounded operator in a Banach space X. The set $\varrho(T)$ of all complex numbers λ such that $(\lambda I - T)^{-1} \in B(\mathrm{X}) = L(\mathrm{X})$ is called the *resolvent set* and the operator $R_\lambda = R_\lambda(T) = (\lambda I - T)^{-1}$ is called the *resolvent*. Let T now be a bounded linear operator in a complex Banach space X. Then the following statements are valid:

a) The set $\varrho(T)$ is open. More precisely, if $\lambda_0 \in \varrho(T)$ and $\lambda \in \mathbb{C}$ satisfy the inequality

$$|\lambda - \lambda_0| < \frac{1}{\|R_{\lambda_0}\|}, \tag{12.155}$$

then R_λ exists and

$$R_\lambda = R_{\lambda_0} + (\lambda - \lambda_0)R^2_{\lambda_0} + (\lambda - \lambda_0)^2 R^3_{\lambda_0} + \ldots = \sum_{k=1}^{\infty}(\lambda - \lambda_0)^{k-1}R^k_{\lambda_0}. \tag{12.156}$$

b) $\{\lambda \in \mathbb{C} : |\lambda| > \|T\|\} \subset \varrho(T)$. More precisely, $\forall\, \lambda \in \mathbb{C}$ with $|\lambda| > \|T\|$, the operator R_λ exists and

$$R_\lambda = -\frac{I}{\lambda} - \frac{T}{\lambda^2} - \frac{T^2}{\lambda^3} - \ldots\,. \tag{12.157}$$

c) $\|R_\lambda - R_{\lambda_0}\| \to 0$, if $\lambda \to \lambda_0 \quad (\lambda, \lambda_0 \in \varrho(T))$, and $\|R_\lambda\| \to 0$, if $\lambda \to \infty \quad (\lambda \in \varrho(T))$.

d) $\left\| \dfrac{R_\lambda - R_{\lambda_0}}{\lambda - \lambda_0} - R^2_{\lambda_0} \right\| \longrightarrow 0$, if $\lambda \to \lambda_0$.

e) For an arbitrary functional $f \in \mathrm{X}^*$ (see 12.5.4.1, p. 681) and arbitrary $x \in \mathrm{X}$ the function $F(\lambda) = f(R_\lambda(x))$ is holomorphic on $\varrho(T)$.

f) For arbitrary $\lambda,\ \mu \in \varrho(T)$, and $\lambda \ne \mu$ one has:

$$R_\lambda R_\mu = R_\mu R_\lambda = \frac{R_\lambda - R_\mu}{\lambda - \mu}. \tag{12.158}$$

12.5.3.2 Spectrum of an Operator

1. Definition of the Spectrum

The set $\sigma(T) = \mathbb{C} \setminus \varrho(T)$ is called the *spectrum* of the operator T. Since $I - T$ has a continuous inverse (and consequently (12.153) has a solution, which continuously depends on the right-hand side) if and only if $1 \in \varrho(T)$, the spectrum $\sigma(T)$ must be known as well as possible. From the properties of the

resolvent set it follows immediately that the spectrum $\sigma(T)$ is a closed set of $\mathbb{C}$ which lies in the disk $\{\lambda \in \mathbb{C}\colon |\lambda| \le \|T\|\}$, however, in many cases $\sigma(T)$ is much smaller than this disk. The spectrum of any linear continuous operator on a complex Banach space is never empty and

$$r(T) = \sup_{\lambda\in\sigma(T)} |\lambda|. \tag{12.159}$$

It is possible to say more about the spectrum in the cases of different special classes of operators. If T is an operator in a finite dimensional space X and if the equation $(\lambda I - T)x = 0$ has only the trivial solution (i.e., $\lambda I - T$ is injective), then $\lambda \in \varrho(T)$ (i.e., $\lambda I - T$ is surjective). If this equation has a non-trivial solution in some Banach space, then the operator $\lambda I - T$ is not injective and $(\lambda I - T)^{-1}$ is in general not defined.

The number $\lambda \in \mathbb{C}$ is called an *eigenvalue* of the linear operator T, if the equation $\lambda x = Tx$ has a non-trivial solution. All those solutions are called *eigenvectors*, or in the case when X is a function space (which occurs very often in applications), they are called *eigenfunctions* of the operator T associated to λ. The subspace spanned by them is called the *eigenspace* (or *characteristic space*) associated to λ. The set $\sigma_p(T)$ of all eigenvalues of T is called the *point spectrum* of the operator T.

2. Comparison to Linear Algebra, Residual Spectrum

An essential difference between the finite dimensional case which is considered in linear algebra and the infinite dimensional case discussed in functional analysis is that in the first case $\sigma(T) = \sigma_p(T)$ always holds, while in the second case the spectrum usually also contains points which are not eigenvalues of T. If $\lambda I - T$ is injective and surjective as well, then $\lambda \in \varrho(T)$ due to the theorem on the continuity of the inverse (see 12.5.2, **7.**, p. 679). In contrast to the finite dimensional case where the surjectivity follows automatically from the injectivity, the infinite dimensional case has to be dealt with in a very different way.

The set $\sigma_c(T)$ of all $\lambda \in \sigma(T)$, for which $\lambda I - T$ is injective and $Im(\lambda I - T)$ is dense in X, is called the *continuous* spectrum and the set $\sigma_r(T)$ of all λ with an injective $\lambda I - T$ and a non-dense image, is called the *residual spectrum* of operator T.

For a bounded linear operator T in a complex Banach space X

$$\sigma(T) = \sigma_p(T) \cup \sigma_c(T) \cup \sigma_r(T), \tag{12.160}$$

where the terms of the right-hand side are mutually disjoint.

12.5.4 Continuous Linear Functionals

12.5.4.1 Definition

For $\mathrm{Y} = \mathbb{F}$ a linear mapping is called a *linear functional* or a *linear form*. In the following discussions, for a Hilbert space the complex case is considered; in other situations almost every times the real case is considered. The Banach space $L(\mathrm{X}, \mathbb{F})$ of all continuous linear functionals is called the *adjoint space* or the *dual space* of X and it is denoted by X^* (sometimes also by X'). The value (in $\mathbb{F}$) of a linear continuous functional $f \in \mathrm{X}^*$ on an element $x \in \mathrm{X}$ is denoted by $f(x)$, often also by (x, f) – emphasizing the bilinear relation of X and X^* – (compare also with the Riesz theorem (see 12.5.4.2, p. 682).

■ **A:** Let $t_1, t_2, \ldots, t_n$ be fixed points of the interval $[a, b]$ and $c_1, c_2, \ldots, c_n$ real numbers. By the formula

$$f(x) = \sum_{k=1}^{n} c_k x(t_k) \tag{12.161}$$

a linear continuous functional is defined on the space $\mathcal{C}([a, b])$; the norm of f is $\|f\| = \sum\limits_{k=1}^{n} |c_k|$. A special case of (12.161) for a fixed $t \in [a, b]$ is the δ functional

$$\delta_t(x) = x(t) \quad (x \in \mathcal{C}([a, b])). \tag{12.162}$$

■ **B:** With an integrable function $\varphi(t)$ (see 12.9.3.1, p. 696) on $[a,b]$

$$f(x) = \int_a^b \varphi(t)x(t)\,dt \qquad (12.163)$$

is a linear continuous functional on $\mathcal{C}([a,b])$ and also on $\mathcal{B}([a,b])$ in each case with the norm $\|f\| = \int_a^b |\varphi(t)|\,dt$.

12.5.4.2 Continuous Linear Functionals in Hilbert Spaces. Riesz Representation Theorem

In a Hilbert space $\mathbb{H}$ equipped with the scalar product $(\cdot,\cdot)$ every element $y \in \mathbb{H}$ defines a linear continuous functional by the formula $f(x) = (x,y)$, where its norm is $\|f\| = \|y\|$. Conversely, if f is a linear continuous functional on $\mathbb{H}$, then there exists a unique element $y \in \mathbb{H}$ such that

$$f(x) = (x,y) \quad \forall\, x \in \mathbb{H}, \qquad (12.164)$$

where $\|f\| = \|y\|$. According to this theorem the spaces $\mathbb{H}$ and $\mathbb{H}^*$ are isomorphic and might be identified.

The Riesz representation theorem contains a hint on how to introduce the notion of *orthogonality in an arbitrary normed space.* Let $A \subset \mathrm{X}$ and $A^* \subset \mathrm{X}^*$. The sets

$$A^\perp = \{f \in \mathrm{X}\colon f(x) = 0 \quad \forall\, x \in A\} \quad \text{and} \quad A^{*\perp} = \{x \in \mathrm{X}\colon f(x) = 0 \quad \forall\, f \in A^*\} \qquad (12.165)$$

are called the *orthogonal complement* or the *annulator* of A and A^*, respectively.

12.5.4.3 Continuous Linear Functionals in L^p

Let $p \geq 1$. The number q is called the *conjugate exponent* to p if $\frac{1}{p} + \frac{1}{q} = 1$, where it is assumed that $q = \infty$ in the case of $p = 1$.

■ Based on the Hölder integral inequality (see 1.4.2.12, p. 32) the functional (12.163) can be considered also in the spaces $L^p([a,b])$ $(1 \leq p \leq \infty)$ (see 12.9.4, p. 697) if $\varphi \in L^q([a,b])$ and $\frac{1}{p} + \frac{1}{q} = 1$. Its norm is then

$$\|f\| = \|\varphi\| = \begin{cases} \left(\displaystyle\int_a^b |\varphi(t)|^q dt\right)^{\frac{1}{q}}, & \text{if} \quad 1 < p \leq \infty, \\ \operatorname*{ess.\,sup}_{t\in[a,b]} |\varphi(t)|, & \text{if} \quad p = 1 \end{cases} \qquad (12.166)$$

(with respect to the definition of ess. sup $|\varphi|$ see (12.221), p. 698). To every linear continuous functional f in the space $L^p([a,b])$ there is a uniquely (up to its equivalence class) defined element $y \in L^q([a,b])$ such that

$$f(x) = (x,y) = \int_a^b x(t)\overline{y(t)}dt, \quad x \in L^p \quad \text{and} \quad \|f\| = \|y\|_q = \left(\int_a^b |y(t)|^q dt\right)^{\frac{1}{q}}. \qquad (12.167)$$

For the case of $p = \infty$ see [12.15].

12.5.5 Extension of a Linear Functional

1. Semi-Norm

A mapping p: $\mathrm{X} \longrightarrow \mathbb{R}$ of a vector space X is called a *semi-norm* or *pseudonorm*, if it has the following properties:

(HN1) $p(x) \geq 0$, (12.168)

(HN2) $p(\alpha x) = |\alpha| p(x)$, (12.169)

(HN3) $p(x+y) \leq p(x) + p(y)$. (12.170)

Comparison with 12.3.1, p. 669, shows that a semi-norm is a norm if and only if $p(x) = 0$ holds only for $x = 0$.

Both for theoretical mathematical questions and for practical reasons in applications of mathematics, the problem of the extension of a linear functional given on a linear subspace $\mathrm{X}_0 \subset \mathrm{X}$ to the entire space (and, in order to avoid trivial and uninteresting cases) with preserving certain " good" properties became a fundamental question. The solution of this problem is guaranteed by

2. Analytic Form of the Hahn-Banach Extension Theorem

Let X be a vector space over $\mathbb{F}$ and p a pseudonorm on X. Let X_0 be a linear (complex in the case of $\mathbb{F} = \mathbb{C}$ and real in the case of $\mathbb{F} = \mathbb{R}$) subspace of X, and let f_0 be a (complex-valued in the case of $\mathbb{F} = \mathbb{C}$ and real-valued in the case of $\mathbb{F} = \mathbb{R}$) linear functional on X_0 satisfying the relation

$$|f_0(x)| \le p(x) \quad \forall\, x \in \mathrm{X}_0. \tag{12.171}$$

Then there exists a linear functional f on X with the following properties:

$$f(x) = f_0(x) \quad \forall\, x \in \mathrm{X}_0, \quad |f(x)| \le p(x) \quad \forall\, x \in \mathrm{X}. \tag{12.172}$$

So, f is an extension of the functional f_0 onto the whole space X preserving the relation (12.171).

If X_0 is a linear subspace of a normed space X and f_0 is a continuous linear functional on X_0, then $p(x) = \|f_0\| \cdot \|x\|$ is a pseudonorm on X satisfying (12.171), so the Hahn-Banach extension theorem for continuous linear functionals is obtained.

Two important consequences are:

1. For every element $x \neq 0$ there is a functional $f \in \mathrm{X}^*$ with $f(x) = \|x\|$ and $\|f\| = 1$.

2. For every linear subspace $\mathrm{X}_0 \subset \mathrm{X}$ and $x_0 \notin \mathrm{X}_0$ with the positive distance $d = \inf_{x \in X_0} \|x - x_0\| > 0$ there is an $f \in \mathrm{X}^*$ such that

$$f(x) = 0 \ \forall\, x \in \mathrm{X}_0, \quad f(x_0) = 1 \quad \text{and} \quad \|f\| = \frac{1}{d}. \tag{12.173}$$

12.5.6 Separation of Convex Sets

1. Hyperplanes

A linear subset L of the real vector space X, $L \neq \mathrm{X}$, is called a *hypersubspace* or *hyperplane through* 0 if there exists an $x_0 \in \mathrm{X}$ such that $\mathrm{X} = lin(x_0, L)$. Sets of the form $x + L$ (L a linear subset) are affine-linear manifolds (see 12.1.2, p. 655). If L is a hypersubspace, these manifolds are called *hyperplanes*.

There exist the following close relations between hypersubspaces, hyperplanes and linear functionals:

a) The kernel $f^{-1}(0) = \{x \in \mathrm{X}\colon f(x) = 0\}$ of a linear functional f on X is a hypersubspace in X, and for each number $\lambda \in \mathbb{R}$ there exists an element $x_\lambda \in \mathrm{X}$ with $f(x_\lambda) = \lambda$ and $f^{-1}(\lambda) = x_\lambda + f^{-1}(0)$.

b) For any given hypersubspace $L \subset X$ and each $x_0 \notin L$ and $\lambda \neq 0$ ($\lambda \in \mathbb{R}$) there always exists a uniquely determined linear functional f on X with $f^{-1}(0) = L$ and $f(x_0) = \lambda$.

The closedness of $f^{-1}(0)$ in the case of a normed space X is equivalent to the continuity of the functional f.

2. Geometric Form of the Hahn–Banach Extension Theorem

Let X be a normed space, $x_0 \in \mathrm{X}$ and L a linear subspace of X. Then for every non-empty convex open set K which does not intersect the affine-linear manifold $x_0 + L$, there exists a closed hypersubspace H such that $x_0 + L \subset H$ and $H \cap K = \emptyset$.

3. Separation of Convex Sets

Two subsets A, B of a real normed space X are called *separated* by a hyperplane if there is a functional $f \in \mathrm{X}^*$ such that:

$$\sup_{x \in A} f(x) \le \inf_{y \in B} f(y). \tag{12.174}$$

The separating hyperplane is then given by $f^{-1}(\alpha)$ with $\alpha = \sup_{x \in A} f(x)$, which means that the two sets are contained in the different half-spaces

$$A \subset \{x \in \mathrm{X}\colon f(x) \le \alpha\} \quad \text{and} \quad B \subset \{x \in \mathrm{X}\colon f(x) \ge \alpha\}. \qquad (12.175)$$

In **Fig. 12.5b,c** two cases of the separation by a hyperplane are shown.
Their disjointness is less decisive for the separation of two sets. In fact, **Fig. 12.5a** shows two sets E and B, which are not separated although E and B are disjoint and B is convex. The convexity of both sets is the intrinsic property for separating them. In this case it is possible that the sets have common points which are contained in the hyperplane.

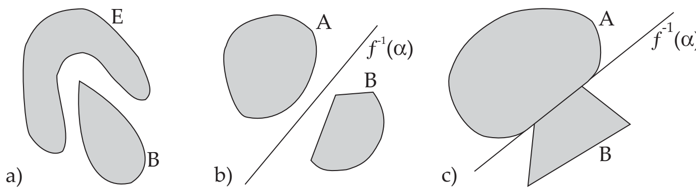


Figure 12.5

If A is a convex set of a normed space X with a non-empty interior $Int(A)$ and $B \subset \mathrm{X}$ is a non-empty convex set with $Int(A) \cap B = \emptyset$, then A and B can be separated. The hypothesis $Int(A) \neq \emptyset$ in that statement cannot be dropped (see [12.3], example 4.47). A (real linear) functional $f \in \mathrm{X}^*$ is called a *supporting functional* of the set A at the point $x_0 \in A$, if there is a real number $\lambda \in \mathbb{R}$ such that $f(x_0) = \lambda$, and $A \subset \{x \in \mathrm{X} : f(x) \le \lambda\}$. $f^{-1}(\lambda)$ is called the *supporting hyperplane* at the point x_0.
For a convex set K with a non-empty interior, there exists a supporting functional at each of its boundary points.

Remark: The famous Kuhn-Tucker theorem (see 18.2, p. 925) which yields practical methods to determine the minimum of convex optimization problems (see [12.5]), is also based on the separation of convex sets.

12.5.7 Second Adjoint Space and Reflexive Spaces

The adjoint space X^* of a normed space X is also a normed space if it is equipped with the norm $\|f\| = \sup\limits_{\|x\|\le 1} |f(x)|$, so $(\mathrm{X}^*)^* = \mathrm{X}^{**}$ – the *second adjoint space* to X can also be considered. The *canonical embedding*

$$J\colon \mathrm{X} \longrightarrow \mathrm{X}^{**} \quad \text{with} \quad Jx = F_x, \text{ where } \quad F_x(f) = f(x) \quad \forall\, f \in \mathrm{X}^* \qquad (12.176)$$

is a norm isomorphism (see 12.3.1, p. 669), hence X is identified with the subset $J(\mathrm{X}) \subset \mathrm{X}^{**}$. A Banach space X is called *reflexive* if $J(\mathrm{X}) = \mathrm{X}^{**}$. Hence the canonical embedding is then a surjective norm isomorphism.

■ Every finite dimensional Banach space and every Hilbert space is reflexive, as well as the spaces L^p $(1 \le p < \infty)$, however $\mathcal{C}([a,b])$, $L^1([0,1])$, $\mathbf{c_0}$ are examples of non-reflexive spaces.

12.6 Adjoint Operators in Normed Spaces

12.6.1 Adjoint of a Bounded Operator

For a given linear continuous operator $T\colon \mathrm{X} \longrightarrow \mathrm{Y}$ (X, Y are normed spaces) to every $g \in \mathrm{Y}^*$ there is assigned a functional $f \in \mathrm{X}^*$ by $f(x) = g(Tx)$, $\forall\, x \in \mathrm{X}$. In this way, a linear continuous operator

$$T^*\colon \mathrm{Y}^* \longrightarrow \mathrm{X}^*, \quad (T^*g)(x) = g(Tx), \quad \forall\, g \in \mathrm{Y}^* \text{ and } \forall\, x \in \mathrm{X} \qquad (12.177)$$

is obtained which is called the *adjoint operator* of T and has the following properties:

$(T+S)^* = T^* + S^*$, $(ST)^* = S^*T^*$, $\|T^*\| = \|T\|$, where for the linear continuous operators $T\colon \mathrm{X} \to \mathrm{Y}$ and $S : \mathrm{Y} \to \mathrm{Z}$ (X, Y, Z normed spaces), the operator $ST : \mathrm{X} \to \mathrm{Z}$ is defined in the natural way

as $ST(x) = S(T(x))$ (see 12.3.4, ■ C, p. 672). With the notation introduced in 12.1.5, p. 658, and 12.5.4.2, p. 682, the following identities are valid for an operator $T \in B(\mathrm{X}, \mathrm{Y})$:

$$\overline{Im(T)} = \ker(T^*)^\perp, \quad \overline{Im(T^*)} = \ker(T)^\perp, \tag{12.178}$$

where the closedness of $Im(T)$ implies the closedness of $Im(T^*)$.

The operator $T^{**}: \mathrm{X}^{**} \to \mathrm{Y}^{**}$, obtained as $(T^*)^*$ from T^*, is called the *second adjoint* of T. Due to $(T^{**}(F_x))g = F_x(T^*g) = (T^*g)(x) = g(Tx) = F_{Tx}(g)$ the operator T^{**} has the following property: If $F_x \in \mathrm{X}^{**}$, then $T^{**}F_x = F_{Tx} \in \mathrm{Y}^{**}$. Hence, the operator $T^{**}: \mathrm{X}^{**} \to \mathrm{Y}^{**}$ is an extension of T.

In a Hilbert space $\mathbb{H}$ the adjoint operator can also be introduced by means of the scalar product $(Tx, y) = (x, T^*y),\ x, y \in \mathbb{H}$. This is based on the Riesz representation theorem, where the identification of $\mathbb{H}$ and $\mathbb{H}^{**}$ implies $(\lambda T)^* = \overline{\lambda} T^*$, $I^* = I$ and even $T^{**} = T$. If T is bijective, then the same holds for T^*, and also $(T^*)^{-1} = (T^{-1})^*$. For the resolvents of T and T^* there holds

$$[R_\lambda(T)]^* = R_{\overline{\lambda}}(T^*), \tag{12.179}$$

from which $\sigma(T^*) = \{\overline{\lambda} \colon \lambda \in \sigma(T)\}$ follows for the spectrum of the adjoint operator.

■ **A:** Let T be an integral operator in the space $L^p([a, b])$ $(1 < p < \infty)$

$$(Tx)(s) = \int_a^b K(s, t) x(t)\, dt \tag{12.180}$$

with a continuous kernel $K(s, t)$. The adjoint operator of T is also an integral operator, namely

$$(T^*g)(t) = \int_a^b K^*(t, s) y_g(s)\, ds \tag{12.181}$$

with the kernel $K^*(s, t) = K(t, s)$, where y_g is the element from L^q associated to $g \in (L^p)^*$ according to (12.167).

■ **B:** In a finite dimensional complex vector space the adjoint of an operator represented by the matrix $\mathbf{A} = (a_{ij})$ is defined by the matrix $\mathbf{A}^*$ with $a^*_{ij} = \overline{a_{ji}}$.

12.6.2 Adjoint Operator of an Unbounded Operator

Let X and Y be real normed spaces and T a (not necessarily bounded) linear operator with a (linear) domain $D(T) \subset \mathrm{X}$ and values in Y. For a given $g \in \mathrm{Y}^*$, the expression $g(Tx)$, depending obviously linearly on x, is meaningful. Now the question is: Does there exist a well-defined functional $f \in \mathrm{X}^*$ such that

$$f(x) = g(Tx) \quad \forall\, x \in D(T). \tag{12.182}$$

Let $D^* \subset \mathrm{Y}^*$ be the set of all those $g \in \mathrm{Y}^*$ for which the representation (12.182) holds for a certain $f \in \mathrm{X}^*$. If $\overline{D(T)} = \mathrm{X}$, then for given g the functional f is uniquely defined. So a linear operator T^* is defined by $f = T^*g$ with $D(T^*) = D^*$. Then for arbitrary $x \in D(T)$ and $g \in D(T^*)$

$$g(Tx) = (T^*g)(x). \tag{12.183}$$

The operator T^* turns out to be closed and is called the *adjoint* of T. The naturalness of this general procedure stems from the fact that $D(T^*) = \mathrm{Y}^*$ holds if and only if T is bounded on $D(T)$. In this case $T^* \in B(\mathrm{Y}^*, \mathrm{X}^*)$ and $\|T^*\| = \|T\|$ hold.

12.6.3 Self-Adjoint Operators

An operator $T \in B(\mathbb{H})$ ($\mathbb{H}$ is a Hilbert space) is called *self-adjoint* if $T^* = T$. In this case

$$(Tx, y) = (x, Ty), \quad x, y \in \mathbb{H} \tag{12.184a}$$

is valid and the number (Tx, x) is real for each $x \in \mathbb{H}$. Then the equality

$$\|T\| = \sup_{\|x\|=1} |(Tx, x)| \tag{12.184b}$$

holds and with $m = m(T) = \inf\limits_{\|x\|=1} (Tx, x)$ and $M = M(T) = \sup\limits_{\|x\|=1} (Tx, x)$ also the relations

$$m(T)\|x\|^2 \le (Tx, x) \le M(T)\|x\|^2 \quad \text{and} \quad \|T\| = r(T) = \max\{|m|, M\} \tag{12.185}$$

are valid. The equality (12.184a) characterizes the self-adjoint operators. The spectrum of a self-adjoint (bounded) operator lies in the interval $[m, M]$ and $m, M \in \sigma(T)$ holds.

12.6.3.1 Positive Definite Operators

A partial ordering can be introduced in the set of all self-adjoint operators of $B(\mathbb{H})$ by defining

$$T \ge 0 \quad \text{if and only if} \quad (Tx, x) \ge 0 \quad \forall\, x \in \mathbb{H}. \tag{12.186}$$

An operator T with $T \ge 0$ is called *positive* (or, more exactly *positive definite*). For any self-adjoint operator T (with **(H1)** from 12.4.1.1, p. 673), $(T^2x, x) = (Tx, Tx) \ge 0$, so T^2 is positive definite. Every positive definite operator T possesses a square root, i.e., there exists a unique positive definite operator W such that $W^2 = T$. Moreover, the vector space of all self-adjoint operators is a K-space (Kantorovich space, see 12.1.7.4, p. 660), where the operators

$$|T| = \sqrt{T^2}, \quad T^+ = \frac{1}{2}(|T| + T), \quad T^- = \frac{1}{2}(|T| - T) \tag{12.187}$$

are the corresponding elements with respect to (12.37). They are of particular importance for the spectral decomposition and spectral and integral representations of self-adjoint operators by means of some Stieltjes integral (see 8.2.3.1, **2.**, p. 506, and [12.1], [12.11], [12.12], [12.15], [12.18]).

12.6.3.2 Projectors in a Hilbert Space

Let $\mathbb{H}_0$ be a subspace of a Hilbert space $\mathbb{H}$. Then every element $x \in \mathbb{H}$ has its projection x' onto $\mathbb{H}_0$ according to the projection theorem (see 12.4.2, p. 674), and therefore, an operator P with $Px = x'$ is defined on $\mathbb{H}$ with values in $\mathbb{H}_0$. P is called a *projector* onto $\mathbb{H}_0$. Obviously, P is linear, continuous, and $\|P\| = 1$. A continuous linear operator P in $\mathbb{H}$ is a projector (onto a certain subspace) if and only if:

a) $P = P^*$, i.e., P is self-adjoint, and

b) $P^2 = P$, i.e., P is *idempotent*.

12.7 Compact Sets and Compact Operators

12.7.1 Compact Subsets of a Normed Space

A subset A of a normed space† X is called

- *compact,* if every sequence of elements from A contains a convergent subsequence whose limit lies in A,
- *relatively compact* or *precompact* if its closure (see 12.2.1.3, p. 664) is compact, i.e., every sequence of elements from A contains a convergent subsequence (whose limit does not necessarily belong to A).

This is the Bolzano–Weierstrass theorem in real calculus for bounded sequences in $\mathbb{R}^n$, and one says that such a set has the *Bolzano–Weierstrass property*.
Every compact set is closed and bounded. Conversely, if the space X is finite dimensional, then every such set is compact. The closed unit ball in a normed space X is compact if and only if X is finite dimensional.
For some characterizations of relatively compact subsets in metric spaces (the Hausdorff theorem on the existence of a finite ε-net) and in the spaces **s**, $\mathcal{C}$ (Arzela–Ascoli theorem) and in the spaces $\boldsymbol{L^p}(1 < p < \infty)$ see [12.15].

12.7.2 Compact Operators

12.7.2.1 Definition of Compact Operator

An arbitrary operator $T\colon \mathrm{X} \longrightarrow \mathrm{Y}$ of a normed space X into a normed space Y is called *compact* if the

†It is enough that X is a metric (or an even more general) space. This generality is not used in what follows.

image $T(A)$ of every bounded set $A \subset \mathrm{X}$ is a relatively compact set in Y. If, in addition the operator T is also continuous, then it is called *completely continuous*. Every *compact linear* operator is bounded and consequently completely continuous. For a linear operator to be compact it is sufficient to require that it transforms the unit ball of X into a relatively compact set in Y.

12.7.2.2 Properties of Linear Compact Operators

A characterization by sequences of the compactness of an operator from $B(\mathrm{X},\mathrm{Y})$ is the following: For every bounded sequence $\{x_n\}_{n=1}^{\infty}$ from X the sequence $\{Tx_n\}_{n=1}^{\infty}$ contains a convergent subsequence. A linear combination of compact operators is also compact. If one of the operators $U \in B(\mathrm{W},\mathrm{X}),\ T \in B(\mathrm{X},\mathrm{Y}),\ S \in B(\mathrm{Y},\mathrm{Z})$ in each of the following products is compact, then the operators TU and ST are also compact. If Y is a Banach space, then the following important statements are valid.

a) Convergence: If a sequence of compact operators $\{T_n\}_{n=1}^{\infty}$ is convergent in the space $B(\mathrm{X},\mathrm{Y})$, then its limit is a compact operator, too.

b) Schauder Theorem: If T is a linear continuous operator, then either both T and T^* are compact or both are not.

c) Spectral Properties of a Compact Operator T in an (Infinite Dimensional) Banach Space X: The zero belongs to the spectrum. Every non-zero point of the spectrum $\sigma(T)$ is an eigenvalue with a finite dimensional eigenspace $\mathrm{X}_\lambda = \{x \in \mathrm{X}\colon (\lambda I - T)x = 0\}$, and $\forall\, \varepsilon > 0$ there is always only a finite number of eigenvalues of T outside the circle $\{|\lambda| \le \varepsilon\}$, where only the zero can be an accumulation point of the set of eigenvalues. If $\lambda = 0$ is not an eigenvalue of T, then T^{-1} is unbounded if it exists.

12.7.2.3 Weak Convergence of Elements

A sequence $\{x_n\}_{n=1}^{\infty}$ of elements of a normed space X is called *weakly convergent* to an element x_0 if for each $f \in \mathrm{X}^*$ the relation $f(x_n) \to f(x_0)$ holds (written as: $x_n \rightharpoonup x_0$ or as $x_n \overset{w}{\to} x_0$).
Obviously: $x_n \to x_0$ implies $x_n \rightharpoonup x_0$. If Y is another normed space and $T\colon \mathrm{X} \longrightarrow \mathrm{Y}$ is a continuous linear operator, then:

a) $x_n \rightharpoonup x_0$ implies $Tx_n \rightharpoonup Tx_0$,

b) if T is compact, then $x_n \rightharpoonup x_0$ implies $Tx_n \to Tx_0$.

■ **A:** Every finite dimensional operator is compact. From this fact it follows that the identity operator in an infinite dimensional space cannot be compact (see 12.7.1, p. 686).

■ **B:** Suppose $\mathrm{X} = \mathbf{l}^2$, and let T be the operator in $\mathbf{l}^2$ given by the infinite matrix

$$\begin{pmatrix} t_{11} & t_{12} & t_{13} & \cdots \\ t_{21} & t_{22} & t_{23} & \cdots \\ t_{31} & \cdot & \cdot & \cdots \\ \cdot & \cdot & \cdot & \cdots \\ \cdot & \cdot & \cdot & \cdots \end{pmatrix} \quad \text{with} \quad Tx = \left(\sum_{k=1}^{\infty} t_{1k}x_k, \ldots, \sum_{k=1}^{\infty} t_{nk}x_k, \ldots \right). \tag{12.188}$$

If $\sum\limits_{k,n=1}^{\infty} |t_{nk}|^2 = M < \infty$, then T is a compact operator from $\mathbf{l}^2$ into $\mathbf{l}^2$ with $\|T\| \le M$.

■ **C:** The integral operator (12.136) is a compact operator in the spaces $\mathcal{C}([a,b])$ and $L^p((a,b))$ $(1 < p < \infty)$.

12.7.3 Fredholm Alternative

Let T be a compact linear operator in a Banach space X. The following equations (of the second kind) are considered with a parameter $\lambda \ne 0$:

$$\begin{aligned} \lambda x - Tx &= y, & \lambda x - Tx &= 0, \\ \lambda f - T^* f &= g, & \lambda f - T^* f &= 0. \end{aligned} \tag{12.189}$$

The following statements are valid:

a) $\dim(\ker(\lambda I - T)) = \dim(\ker(\lambda I - T^*)) < +\infty$, i.e., both homogeneous equations always have the same number of linearly independent solutions.

b) $\mathrm{Im}(\lambda I - T) = \ker(\lambda I - T^*)^{\perp}$ and [‡] $\mathrm{Im}(\lambda I - T^*) = \ker(\lambda I - T)^{\perp}$.

c) $\mathrm{Im}(\lambda I - T) = \mathrm{X}$ if and only if $\ker(\lambda I - T) = 0$.

d) The *Fredholm alternative* (also called the Riesz–Schauder theorem):

α) Either the homogeneous equation has only the trivial solution. In this case $\lambda \in \varrho(T)$, the operator $(\lambda I - T)^{-1}$ is bounded, and the inhomogeneous equation has exactly one solution $x = (\lambda I - T)^{-1}y$ for arbitrary $y \in \mathrm{X}$.

β) Or the homogeneous equation has at least one non-trivial solution. In this case λ is an eigenvalue of T, i.e., $\lambda \in \sigma(T)$, and the inhomogeneous equation has a (non-unique) solution if and only if the right-hand side y satisfies the condition $f(y) = 0$ for every solution f of the adjoint equation $T^*f = \lambda f$. In this last case every solution x of the inhomogeneous equation has the form $x = x_0 + h$, where x_0 is a fixed solution of the inhomogeneous equation and $h \in \ker(\lambda I - T)$.

Linear equations of the form $Tx = y$ with a compact operator T are called equations of the first kind. Their mathematical investigation is in general more difficult (see [12.11],[12.18]).

12.7.4 Compact Operators in Hilbert Space

Let $T : \mathbb{H} \longrightarrow \mathbb{H}$ be a compact operator. Then T is the limit (in $B(\mathbb{H})$) of a sequence of finite dimensional operators. The similarity to the finite dimensional case can be seen from the statements:

If C is a finite dimensional operator and $T = I - C$, then the injectivity of T implies the existence of T^{-1} and $T^{-1} \in B(\mathbb{H})$.

If C is a compact operator, then the following statements are equivalent:

a) $\exists T^{-1}$ and it is continuous,

b) $x \neq 0 \Rightarrow Tx \neq 0$, i.e., T is injective,

c) $T(\mathbb{H}) = \mathbb{H}$, i.e., T is surjective.

12.7.5 Compact Self-Adjoint Operators

1. Eigenvalues

A compact self-adjoint operator $T \neq 0$ in a Hilbert space $\mathbb{H}$ possesses at least one non-zero eigenvalue. More precisely, T always has an eigenvalue λ with $|\lambda| = \|T\|$. The set of eigenvalues of T is at most countable.

Any compact self-adjoint operator T has the representation $T = \sum_k \lambda_k P_{\lambda_k}$ (in $B(\mathbb{H})$), where λ_k are the different eigenvalues of T and P_λ denotes the projector onto the eigenspace $\mathbb{H}_\lambda$. In this case the operator T is diagonalizable. From this fact it follows that $Tx = \sum_k \lambda_k (x, e_k)e_k$ for every $x \in \mathbb{H}$, where $\{e_k\}$ is the orthonormal system of the eigenvectors of T. If $\lambda \notin \sigma(T)$ and $y \in \mathbb{H}$, then the solution of the equation $(\lambda I - T)x = y$ can be represented as $x = R_\lambda(T)y = \sum_k \frac{1}{\lambda - \lambda_k}(y, e_k)e_k$.

2. Hilbert-Schmidt Theorem

If T is a compact self-adjoint operator in a separable Hilbert space $\mathbb{H}$, then there is a basis in $\mathbb{H}$ consisting of the eigenvectors of T.

The so-called spectral (mapping) theorems (see [12.8], [12.10], [12.12], [12.13], [12.18]) can be considered as the generalization of the Hilbert–Schmidt theorem for the non-compact case of self-adjoint (bounded or unbounded) operators.

[‡] Here the orthogonality is considered in Banach spaces (see 12.5.4.2, p. 682).

12.8 Non-Linear Operators

In the theory of non-linear operator equations the most important methods are based on the following principles:

1. Principle of the Contracting Mapping, Banach Fixed-Point Theorem (see 12.2.2.3, p. 666, and 12.2.2.4, p. 666). For further modifications of this principle see [12.8],[12.11], [12.12], [12.18].

2. Generalization of the Newton Method (see 18.2.5.2, p. 931 and 19.1.1.2, p. 950) for the infinite dimensional case.

3. Schauder Fixed-Point Principle (see 12.8.4, p. 691)

4. Leray-Schauder Theory (see 12.8.5, p. 692)

Methods based on principles **1** and **2** yield information on the existence, uniqueness, constructivity etc. of the solution, while methods based on principles **3** and **4**, in general, allow "only" the qualitative statement of the existence of a solution. If further properties of operators are known then see also 12.8.6, p. 692, and 12.8.7, p. 693.

12.8.1 Examples of Non-Linear Operators

For non-linear operators the relation between continuity and boundedness discussed for linear operators in 12.5.1, p. 677 is no longer valid in general. In studying non-linear operator equations, e.g., non-linear boundary value problems or integral equations, the following non-linear operators occur most often. Iteration methods described in 12.2.2.4, p. 666, can be successfully applied for solving non-linear integral equations.

1. Nemytskij Operator

Let Ω be an open measurable subset from $\mathbb{R}^n$ (12.9.1, p. 693) and $f\colon\ \Omega\times\mathbb{R}\longrightarrow\mathbb{R}$ a function of two variables $f(x,s)$, which is continuous with respect to x for almost every s and measurable with respect to s for every x (*Caratheodory conditions*). The non-linear operator $\mathcal{N}$ to $\mathcal{F}(\Omega)$ defined as

$$(\mathcal{N}u)(x) = f[x, u(x)] \quad (x \in \Omega) \tag{12.190}$$

is called the *Nemytskij operator*. It is continuous and bounded if it maps $L^p(\Omega)$ into $L^q(\Omega)$, where $\frac{1}{p}+\frac{1}{q}=1$. This is the case, e.g., if

$$|f(x,s)| \le a(x) + b|s|^{\frac{p}{q}} \quad \text{with} \quad a(x) \in L^q(\Omega) \quad (b > 0) \tag{12.191}$$

or if $f\colon\ \Omega\times\mathbb{R}\longrightarrow\mathbb{R}$ is continuous. The operator $\mathcal{N}$ is compact only in special cases.

2. Hammerstein Operator

Let Ω be a relatively compact subset of $\mathbb{R}^n$, f a function satisfying the Caratheodory conditions and $K(x,y)$ a continuous function on $\overline{\Omega}\times\overline{\Omega}$. The non-linear operator $\mathcal{H}$ on $\mathcal{F}(\Omega)$

$$(\mathcal{H}u)(x) = \int_{\Omega} K(x,y) f[y, u(y)]\, dy \quad (x \in \Omega) \tag{12.192}$$

is called the *Hammerstein operator*. $\mathcal{H}$ can be written in the form $\mathcal{H}=\mathcal{K}\cdot\mathcal{N}$ with the Nemytskij operator $\mathcal{N}$ and the integral operator $\mathcal{K}$ determined by the kernel K

$$(\mathcal{K}u)(x) = \int_{\Omega} K(x,y) u(y)\, dy \quad (x \in \Omega). \tag{12.193}$$

If the kernel $K(x,y)$ satisfies the additional condition

$$\int_{\Omega\times\Omega} |K(x,y)|^q\, dx\, dy < \infty \tag{12.194}$$

and the function f satisfies the condition (12.191), then $\mathcal{H}$ is a continuous and compact operator on $L^p(\Omega)$.

3. Urysohn Operator

Let $\Omega \subset \mathbb{R}^n$ be an open measurable subset and $K(x, y, s)\colon\ \Omega \times \Omega \times \mathbb{R} \longrightarrow \mathbb{R}$ a function of three variables. Then the non-linear operator $\mathcal{U}$ on $\mathcal{F}(\Omega)$

$$(\mathcal{U}u)(x) = \int_\Omega K[x, y, u(y)]\, dy \quad (x \in \Omega) \tag{12.195}$$

is called the *Urysohn operator*. If the kernel K satisfies the appropriate conditions, then $\mathcal{U}$ is a continuous and compact operator in $\mathcal{C}(\Omega)$ or in $L^p(\Omega)$, respectively.

12.8.2 Differentiability of Non-Linear Operators

Let X, Y be Banach spaces, $D \subset$ X be an open set and $T\colon D \longrightarrow$ Y. The operator T is called *Fréchet differentiable* (or, briefly, differentiable) at the point $x \in D$ if there exists a linear operator $L \in B(\mathrm{X}, \mathrm{Y})$ (in general depending on the point x) such that

$$T(x + h) - T(x) = Lh + \omega(h) \quad \text{with} \quad \|\omega(h)\| = o(\|h\|) \tag{12.196}$$

or in an equivalent form

$$\lim_{\|h\| \to 0} \frac{\|T(x + h) - T(x) - Lh\|}{\|h\|} = 0, \tag{12.197}$$

i.e., $\forall\, \varepsilon > 0,\ \exists\, \delta > 0$, such that $\|h\| < \delta$ implies $\|T(x + h) - T(x) - Lh\| \le \varepsilon\|h\|$. The operator L, which is usually denoted by $T'(x)$, $T'(x, \cdot)$ or $T'(x)(\cdot)$, is called the *Fréchet derivative* of the operator T at the point x. The value $dT(x; h) = T'(x)h$ is called the *Fréchet differential* of the operator T at the point x (for the increment h).

The differentiability of an operator at a point implies its continuity at that point. If $T \in B(\mathrm{X}, \mathrm{Y})$, i,e., T itself is linear and continuous, then T is differentiable at every point, and its derivative is equal to T.

12.8.3 Newton's Method

Let X, D be as in the previous paragraph and $T\colon D \longrightarrow$ Y. Under the assumption of the differentiability of T at every point of the set D an operator $T'\colon D \longrightarrow B(\mathrm{X}, \mathrm{Y})$ is defined by assigning the element $T'(x) \in B(\mathrm{X}, \mathrm{Y})$ to every point $x \in D$. Suppose the operator T' is continuous on D (in the operator norm); in this case T is called *continuously differentiable* on D.

Suppose Y = X and also that the set D contains a solution x^* of the equation

$$T(x) = 0. \tag{12.198}$$

Furthermore, it is assumed that the operator $T'(x)$ is continuously invertible for each $x \in D$, hence $[T'(x)]^{-1}$ is in $B(\mathrm{X})$. Because of (12.196) for an arbitrary $x_0 \in D$ one conjectures that the elements $T(x_0) = T(x_0) - T(x^*)$ and $T'(x_0)(x_0 - x^*)$ are "not far" from each other and therefore the element x_1 defined as

$$x_1 = x_0 - [T'(x_0)]^{-1}T(x_0) \tag{12.199}$$

is an approximation of x^* (under the given assumptions). Starting with an arbitrary x_0 the so-called *Newton approximation sequence*

$$x_{n+1} = x_n - [T'(x_n)]^{-1}T(x_n) \quad (n = 0, 1, \ldots) \tag{12.200}$$

can be constructed. There are many theorems known from the literature discussing the behavior and the convergence properties of this method. Here only the following most important result is mentioned which demonstrates the main properties and advantages of Newton's method:

$\forall\, \varepsilon \in (0, 1)$ there exists a ball $B = B(x_0; \delta),\ \delta = \delta(\varepsilon)$ in X, such that all points x_n lie in B and the Newton sequence converges to the solution x^* of (12.198). Moreover, $\|x_n - x_0\| \le \varepsilon^n\|x_0 - x^*\|$ which yields a practical error estimation.

The *modified Newton's method* is obtained if the operator $[T'(x_0)]^{-1}$ is used instead of $[T'(x_n)]^{-1}\ \forall n = 0, 1, \ldots$ in formula (12.200). For further estimations of the speed of convergence and for the (in general sensitive) dependence of the method on the choice of the starting point x_0 see [12.7], [12.12], [12.18].

■ **Jacobian or Functional Matrix** Given a non-linear operator $T = F\colon D \longrightarrow \mathbb{R}^m$ on an open set $D \subset \mathbb{R}^n$ with m non-linear coordinate functions $F_1, F_2, \ldots, F_m$ and n independent variables $x_1, x_2, \ldots, x_n$. Then

$$F(x) = \begin{pmatrix} F_1(x) \\ F_2(x) \\ \vdots \\ F_m(x) \end{pmatrix} \in \mathbb{R}^m \quad \forall\, x = (x_1, x_2, \ldots, x_n) \in D \tag{12.201}$$

holds. If the partial derivatives $\dfrac{\partial F_i}{\partial x_k}$ $(k = 1, 2, \ldots, n)$ of the coordinate functions F_i $(i = 1, 2, \ldots, m)$ on D exist and are continuous, then the mapping (the operator) F in every point of D is differentiable and its derivative at the point $x = (x_1, x_2, \ldots, x_n) \in D$ is the linear operator $F'(x) : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ with the matrix representation

$$F'(x) = \begin{pmatrix} \dfrac{\partial F_1(x)}{\partial x_1} & \dfrac{\partial F_1(x)}{\partial x_2} & \cdots & \dfrac{\partial F_1(x)}{\partial x_n} \\ \dfrac{\partial F_2(x)}{\partial x_1} & \dfrac{\partial F_2(x)}{\partial x_2} & \cdots & \dfrac{\partial F_2(x)}{\partial x_n} \\ \vdots & \vdots & \vdots & \vdots \\ \dfrac{\partial F_m(x)}{\partial x_1} & \dfrac{\partial F_m(x)}{\partial x_2} & \cdots & \dfrac{\partial F_m(x)}{\partial x_n} \end{pmatrix}. \tag{12.202}$$

The derivative $F'(x)$ is a matrix of the type (m, n). It is called the *Jacobian* or *functional matrix* of F. Special cases of occurrence are, e.g., the iterative solution of systems of non-linear equations by the Newton method (see 19.2.2.2, p. 962) or describing the independence of functions (see 2.18.2.6, **3.**, p. 123).
For $m = n$ the so-called *functional determinant* or *Jacobian determinant* can be formed, which is denoted shortly by

$$\frac{D(F_1, F_2, \ldots, F_n)}{D(x_1, x_2, \ldots, x_n)}. \tag{12.203}$$

This determinant is used for the solution of (mostly inner-mathematical) problems (see also, e.g., 8.5.3.2, p. 539).

12.8.4 Schauder's Fixed-Point Theorem

Let $T : D \longrightarrow X$ be a non-linear operator defined on a subset D of a Banach space X. The non-trivial question of whether the equation $x = T(x)$ has at least one solution, can be answered as follows: If $\mathrm{X} = \mathbb{R}$ and $D = [-1, 1]$, then every continuous function, mapping D into D, has a fixed point in D. If X is an arbitrary *finite dimensional* normed space (dimX ≥ 2), then *Brouwer's fixed-point theorem* holds.

1. Brouwer's Fixed-Point Theorem Let D be a non-empty closed bounded and convex subset of a finite dimensional normed space. If T is a continuous operator, which maps D into itself, then T has at least one fixed point in D.
The answer in the case of an arbitrary infinite dimensional Banach space X is given by *Schauder's fixed-point theorem.*

2. Schauder's Fixed-Point Theorem Let D be a non-empty closed bounded and convex subset of a Banach space X. If the operator $T : D \longrightarrow \mathrm{X}$ is continuous and compact (hence completely continuous) and it maps D into itself, then T has at least one fixed point in D.
By using this theorem, it is proved, e.g., that the initial value problem (12.70), p. 668, always has a

local solution for $t \geq 0$, if the right-hand side is assumed only to be continuous.

12.8.5 Leray-Schauder Theory

For the existence of solutions of the equations $x = T(x)$ and $(I+T)(x) = y$ with a completely continuous operator T, a further principle is found which is based on deep properties of the mapping degree. It can be successfully applied to prove the existence of a solution of non-linear boundary value problems. Here only those results of this theory are mentioned which are the most useful ones in practical problems, and for simplicity a formulation is chosen which avoids the notion of the mapping degree.

Leray–Schauder Theorem: Let D be an open bounded set in a real Banach space X and let $T : \overline{D} :\longrightarrow$ X be a completely continuous operator. Let $y \in D$ be a point such that $x + \lambda T(x) \neq y$ for each $x \in \partial D$ and $\lambda \in [0, 1]$, where ∂D denotes the boundary of the set D. Then the equation $(I+T)(x) = y$ has at least one solution.

The following version of this theorem is very useful in applications:

Let T be a completely continuous operator in the Banach space X. If all solutions of the family of equations

$$x = \lambda T(x) \quad (\lambda \in [0, 1]) \tag{12.204}$$

are uniformly bounded, i.e., $\exists\, c > 0$ such that $\forall\, \lambda$ and $\forall\, x$ satisfying (12.204) the a priori estimation $\|x\| \leq c$ holds, then the equation $x = T(x)$ has a solution.

12.8.6 Positive Non-Linear Operators

The successful application of Schauder's fixed-point theorem requires the choice of a set with appropriate properties, which is mapped into itself by the considered operator. In applications, especially in the theory of non-linear boundary value problems, ordered normed function spaces and positive operators are often considered, i.e., which leave the corresponding cone invariant, or *isotone increasing* operators, i.e., if $x \leq y \Rightarrow T(x) \leq T(y)$. If confusions (see, e.g., 12.8.7, p. 693) are excluded, these operators are also called *monotone*.

Let X $=$ (X, X_+, $\|\cdot\|$) be an ordered Banach space, X_+ a closed cone and $[a, b]$ an order interval of X. If X_+ is normal and T is a completely continuous (not necessarily isotone) operator that satisfies $T([a, b]) \subset [a, b]$, then T has at least one fixed point in $[a, b]$ **(Fig. 12.6b)**.

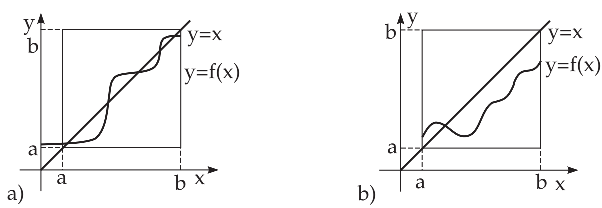


Figure 12.6

Notice that the condition $T([a, b]) \subset [a, b]$ automatically holds for any isotone increasing operator T, which is defined on an (o)-interval (order interval) $[a, b]$ of the space X if it maps only the endpoints a, b into $[a, b]$, i.e., when the two conditions $T(a) \geq a$ and $T(b) \leq b$ are satisfied. Then both sequences

$$x_0 = a \text{ and } x_{n+1} = T(x_n) \;\; (n \geq 0) \quad \text{and} \quad y_0 = b \text{ and } y_{n+1} = T(y_n) \;\; (n \geq 0) \tag{12.205}$$

are well defined, i.e., $x_n, y_n \in [a, b],\ n \geq 0$. They are monotone increasing and decreasing, respectively, i.e., $a = x_0 \leq x_1 \leq \ldots \leq x_n \leq \ldots$ and $b = y_0 \geq y_1 \geq \ldots y_n \geq \ldots$. A fixed point x_*, x^* of the operator T is called *minimal*, *maximal*, respectively, if for every fixed point z of T the inequalities $x_* \leq z$, $z \leq x^*$ hold, respectively.

Now,the following statement is valid (**Fig. 12.6a)**): Let X be an ordered Banach space with a closed

cone X_+ and $T\colon D \longrightarrow X$, $D \subset X$ a continuous isotone increasing operator. Let $[a,b] \subset D$ be such that $T(a) \geq a$ and $T(b) \leq b$. Then $T([a,b]) \subset [a,b]$, and the operator T has a fixed point in $[a,b]$ if one of the following conditions is fulfilled:

a) X_+ is normal and T is compact;

b) X_+ is regular.

Then the sequences $\{x_n\}_{n=0}^{\infty}$ and $\{y_n\}_{n=0}^{\infty}$, defined in (12.205), converge to the minimal and to the maximal fixed points of T in $[a,b]$, respectively.

The notion of the *super- and sub-solutions* is based on these results (see [12.14]).

12.8.7 Monotone Operators in Banach Spaces

1. Special Properties

An arbitrary operator $T\colon D \subset X \longrightarrow Y$ (X, Y normed spaces) is called *demi-continuous* at the point $x_0 \in D$ if for each sequence $\{x_n\}_{n=1}^{\infty} \subset D$ converging to x_0 (in the norm of X) the sequence $\{T(x_n)\}_{n=1}^{\infty}$ converges weakly to $T(x_0)$ in Y. T is called *demi-continuous* on the set D if T is *demi-continuous* at every point of D.

In this paragraph another generalization of the notion of monotonicity known from real analysis are introduced. Let X now be a real Banach space, X^* its dual, $D \subset X$ and $T\colon D \longrightarrow X^*$ a non-linear operator. T is called *monotone* if $\forall\, x, y \in D$ the inequality $(T(x) - T(y), x - y) \geq 0$ holds. If $X = \mathbb{H}$ is a Hilbert space, then $(\cdot,\cdot)$ means the scalar product, while in the case of an arbitrary Banach space one refers to the notation introduced in 12.5.4.1, p. 681. The operator T is called *strongly monotone* if there is a constant $c > 0$ such that $(T(x) - T(y), x - y) \geq c\|x - y\|^2$ for $\forall\, x, y \in D$. An operator $T\colon X \longrightarrow X^*$ is called *coercive* if $\lim\limits_{\|x\|\to\infty} \dfrac{(T(x), x)}{\|x\|} = \infty$.

2. Existence Theorems

for solutions of operator equations with monotone operators are given here only exemplarily: If the operator T, mapping the real separable Banach space X into X^*, ($D_T = X$), is monotone demi-continuous and coercive, then the equation $T(x) = f$ has a solution for arbitrary $f \in X^*$.

If in addition the operator T is strongly monotone, then the solution is unique. In this case the inverse operator T^{-1} also exists.

For a monotone, demi-continuous operator $T\colon \mathbb{H} \longrightarrow \mathbb{H}$ in a Hilbert space $\mathbb{H}$ with $D_T = \mathbb{H}$, there holds $Im(I + T) = \mathbb{H}$, where $(I + T)^{-1}$ is continuous. If T is supposed to be strongly monotone, then T^{-1} is bijective with a continuous T^{-1}.

Constructive approximation methods for the solution of the equation $T(x) = 0$ with a monotone operator T in a Hilbert space are based on the idea of Galerkin's method (see 19.4.2.2, p. 974, or [12.10], [12.18]). By means of this theory set-valued operators $T\colon X \longrightarrow 2^{X^*}$ can also be handled. The notion of monotonicity is then generalized by $(f - g, x - y) \geq 0$, $\forall\, x, y \in D_T$ and $f \in T(x), g \in T(y)$.

12.9 Measure and Lebesgue Integral

12.9.1 Set Algebras and Measures

The initial point for introducing measures is a generalization of the notion of the length of an interval in $\mathbb{R}$, of the area, and of the volume of subsets of $\mathbb{R}^2$ and $\mathbb{R}^3$, respectively. This generalization is necessary in order to "measure" as many sets as possible and to "make integrable" as many functions as possible. For instance, the volume of an n-dimensional rectangular parallelepiped

$$Q = \{x \in \mathbb{R}^n\colon\ a_k \leq x_k \leq b_k \quad (k = 1, 2, \ldots, n)\} \quad \text{has the value} \quad \prod_{k=1}^{n} (b_k - a_k). \tag{12.206}$$

1. σ Algebra or Set Algebra

Let X be an arbitrary set. A non-empty system $\mathcal{A}$ of subsets from X is called a *σ algebra* if:

a) $A \in \mathcal{A}$ implies $X \setminus A \in \mathcal{A}$ and (12.207a)

b) $A_1, A_2, \ldots, A_n, \ldots \in \mathcal{A}$ implies $\bigcup_{n=1}^{\infty} A_n \in \mathcal{A}$. (12.207b)

Every σ algebra contains the sets $\emptyset$ and X, the intersection of countably many of its sets and also the difference sets of any two of its sets.

In the following $\overline{\mathbb{R}}$ denotes the set $\mathbb{R}$ of real numbers extended by the elements $-\infty$ and $+\infty$ (extended real line), where the algebraic operations and the order properties from $\mathbb{R}$ are extended to $\overline{\mathbb{R}}$ in the natural way. The expressions $(\pm\infty) + (\mp\infty)$ and $\frac{\infty}{\infty}$ are meaningless, while $0 \cdot (+\infty)$ and $0 \cdot (-\infty)$ are assigned the value 0.

2. Measure

A function μ: $\mathcal{A} \longrightarrow \overline{\mathbb{R}}_+ = \mathbb{R}_+ \cup +\infty$, defined on a σ algebra $\mathcal{A}$, is called a *measure* if

a) $\mu(A) \geq 0 \quad \forall\, A \in \mathcal{A}$, (12.208a)

b) $\mu(\emptyset) = 0$, (12.208b)

c) $A_1, A_2, \ldots A_n, \ldots \in \mathcal{A},\ A_k \cap A_l = \emptyset\ (k \neq l)$ implies $\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mu(A_n)$. (12.208c)

The property c) is called *σ additivity* of the measure. If μ is a measure on $\mathcal{A}$, and for the sets $A, B \in \mathcal{A}$, $A \subset B$ holds, then $\mu(A) \leq \mu(B)$ (*monotonicity*). If $A_n \in \mathcal{A}$ $(n = 1, 2, \ldots)$ and $A_1 \subset A_2 \subset \cdots$, then $\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \lim_{n\to\infty} \mu(A_n)$ (*continuity from below*).

Let $\mathcal{A}$ be a σ algebra of subsets of X and μ a measure on $\mathcal{A}$. The triplet $X = (X, \mathcal{A}, \mu)$ is called a *measure space*, and the sets belonging to $\mathcal{A}$ are called *measurable* or *$\mathcal{A}$-measurable*.

■ **A: Counting Measure:** Let X be a finite set $\{x_1, x_2, \ldots, x_N\}$, $\mathcal{A}$ the σ algebra of all subsets of X, and let assign a non-negative number p_k to each x_k $(k = 1, \ldots, N)$. Then the function μ defined on $\mathcal{A}$ for every set $A \in \mathcal{A}$, $A = \{x_{n_1}, x_{n_2}, \ldots, x_{n_k}\}$ by $\mu(A) = p_{n_1} + p_{n_2} + \cdots + p_{n_k}$ is a measure which takes on only finite values since $\mu(X) = p_1 + \cdots + p_N < \infty$. This measure is called the *counting measure*.

■ **B: Dirac Measure:** Let $\mathcal{A}$ be a σ algebra of subsets of a set X and a an arbitrary given point from X. Then a measure (called *Dirac Measure*) is defined on $\mathcal{A}$ by

$$\delta_a(A) = \begin{cases} 1, & \text{if } a \in A, \\ 0, & \text{if } a \notin A. \end{cases} \tag{12.209a}$$

It is called the *δ function* (concentrated on a). The *characteristic function* or *indicator function* of a subset $A \subseteq X$ denotes the function χ_A: $X \longrightarrow \{0, 1\}$ of X on $\{0, 1\}$, which has the value 1 for $x \in A$ and for all other x the value 0:

$$\chi_A(x) = \begin{cases} 1, & \text{if } x \in A, \\ 0, & \text{otherwise.} \end{cases} \tag{12.209b}$$

Obviously $\delta_a(A) = \delta_a(\chi_A) = \chi_A(a)$ (see 12.5.4, p. 681), where χ_A denotes the characteristic function of the set A.

■ **C: Lebesgue Measure:** Let X be a metric space and $\mathcal{B}(X)$ the smallest σ algebra of subsets of X which contains all the open sets from X. $\mathcal{B}(X)$ exists as the intersection of all the σ algebras containing all the open sets, and is called the *Borel σ algebra* of X. Every element from $\mathcal{B}(X)$ is called a *Borel set* (see [12.6]).

Suppose now, $X = \mathbb{R}^n$ $(n \geq 1)$. Using an extension procedure a σ algebra and a measure on it can be constructed, which coincides with the volume on the set of all rectangular parallelepipeds in $\mathbb{R}^n$. More precisely: There exists a uniquely defined σ algebra $\mathcal{A}$ of subsets of $\mathbb{R}^n$ and a uniquely defined measure λ on $\mathcal{A}$ with the following properties:

a) Each open set from $\mathbb{R}^n$ belongs to $\mathcal{A}$, in other words: $\mathcal{B}(\mathbb{R}^n) \subset \mathcal{A}$.

b) If $A \in \mathcal{A}$, $\lambda(A) = 0$ and $B \subset A$ then $B \in \mathcal{A}$ and $\lambda(B) = 0$.

c) If Q is a rectangular parallelepiped, then $Q \in \mathcal{A}$, and $\lambda(Q) = \prod_{k=1}^{n} (b_k - a_k)$.

d) λ is translation invariant, i.e., for every vector $x \in \mathbb{R}^n$ and every set $A \in \mathcal{A}$ one has $x + A = \{x + y\colon y \in A\} \in \mathcal{A}$ and $\lambda(x + A) = \lambda(A)$.

The elements of $\mathcal{A}$ are called *Lebesgue measurable* subsets of $\mathbb{R}^n$. λ is the (n-dimensional) *Lebesgue measure* in $\mathbb{R}^n$.

Remark: In measure theory and integration theory one says that a certain statement (property, or condition) with respect to the measure μ is valid *almost everywhere* or *μ-almost everywhere* on a set X, if the set, where the statement is not valid, has measure zero. It is denoted by a.e. or μ-a.e.[§] For instance, if λ is the Lebesgue measure on $\mathbb{R}$ and A, B are two disjoint sets with $\mathbb{R} = A \cup B$ and f is a function on $\mathbb{R}$ with $f(x) = 1,\ \forall x \in A$ and $f(x) = 0,\ \forall\, x \in B$, then $f = 1,\ \lambda$-a.e. on $\mathbb{R}$ if and only if $\lambda(B) = 0$.

12.9.2 Measurable Functions

12.9.2.1 Measurable Function

Let $\mathcal{A}$ be a σ algebra of subsets of a set X. A function $f\colon \mathrm{X} \longrightarrow \overline{\mathbb{R}}$ is called *measurable* if for an arbitrary $\alpha \in \mathbb{R}$ the set $f^{-1}((\alpha, +\infty]) = \{x : x \in \mathrm{X},\ f(x) > \alpha\}$ is in $\mathcal{A}$.
A complex-valued function $g + \mathrm{i}h$ is called measurable if both functions g and h are measurable. The characteristic function χ_A of every set $A \in \mathcal{A}$ is measurable, because

$$\chi_A^{-1}((\alpha, +\infty]) = \begin{cases} A, & \text{if } \alpha \in (-\infty, 1), \\ \emptyset, & \text{if } \alpha \geq 1 \end{cases} \tag{12.210}$$

is valid (see Dirac measure, p. 694). If $\mathcal{A}$ is the σ algebra of the Lebesgue measurable sets of $\mathbb{R}^n$ and $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ is a continuous function, then the set $f^{-1}((\alpha, +\infty]) = f^{-1}((\alpha, +\infty))$, according to 12.2.3, p. 668, is open for every $\alpha \in \mathbb{R}$, hence f is measurable.

12.9.2.2 Properties of the Class of Measurable Functions

The notion of measurable functions requires no measure but a σ algebra. Let $\mathcal{A}$ be a σ algebra of subsets of the set X and let $f, g, f_n : \mathrm{X} \longrightarrow \overline{\mathbb{R}}$ be measurable functions. Then the following functions (see 12.1.7.4, p. 660) are also measurable:

a) αf for every $\alpha \in \mathbb{R}$; $f \cdot g$;

b) $f_+,\ f_-,\ |f|,\ f \vee g$ and $f \wedge g$;

c) $f + g$, if there is no point from X where the expression $(\pm\infty) + (\mp\infty)$ occurs;

d) $\sup f_n,\ \inf f_n,\ \limsup f_n\ (= \lim\limits_{n\to\infty} \sup\limits_{k\geq n} f_k),\ \liminf f_n$;

e) the point-wise limit $\lim f_n$, in case it exists;

f) if $f \geq 0$ and $p \in \mathbb{R}, p > 0$, the f^p is measurable.

A function $f\colon \mathrm{X} \longrightarrow \mathbb{R}$ is called *elementary* or *simple* if there is a finite number of pairwise disjoint sets $A_1, \ldots, A_n \in \mathcal{A}$ and real numbers $\alpha_1, \ldots, \alpha_n$ such that $f = \sum_{k=1}^{n} \alpha_k \chi_k$, where χ_k denotes the characteristic function of the set A_k. Since each characteristic function of a measurable set is measurable (see (12.210)), so every elementary function is measurable. It is interesting that each measurable function can be approximated arbitrarily well by elementary functions: For each measurable function $f \geq 0$ there exists a monotone increasing sequence of non-negative elementary functions, which converges point-wise to f.

[§] Here and in the following parts "a.e." is an abbreviation for "almost everywhere".

12.9.3 Integration

12.9.3.1 Definition of the Integral

Let $(\mathrm{X}, \mathcal{A}, \mu)$ be a measure space. The integral $\int_X f\,d\mu$ (also denoted by $\int f\,d\mu$) for a measurable function f is defined by means of the following steps:

1. If f is an elementary function $f = \sum\limits_{k=1}^{n} \alpha_k \chi_k$, then

$$\int f\,d\mu = \sum_{k=1}^{n} \alpha_k \mu(A_k). \tag{12.211}$$

2. If $f\colon \mathrm{X} \longrightarrow \overline{\mathbb{R}}$ $(f \geq 0)$, then

$$\int f\,d\mu = \sup\left\{\int g\,d\mu : \quad g \text{ is an elementary function with } 0 \leq g(x) \leq f(x),\ \forall\, x \in \mathrm{X}\right\}. \tag{12.212}$$

3. If $f\colon \mathrm{X} \longrightarrow \overline{\mathbb{R}}$ and f_+, f_- are the positive and the negative parts of f, then

$$\int f\,d\mu = \int f_+\,d\mu - \int f_-\,d\mu \tag{12.213}$$

under the condition that at least one of the integrals on the right side is finite (in order to avoid the meaningless expression $\infty - \infty$).

4. For a complex-valued function $f = g + \mathrm{i}h$, if the integrals (12.213) of the functions g, h are finite, put

$$\int f\,d\mu = \int g\,d\mu + \mathrm{i}\int h\,d\mu. \tag{12.214}$$

5. If for any measurable set A and a function f there exists the integral of the function $f\chi_A$ then put

$$\int\limits_A f\,d\mu := \int f\chi_A\,d\mu. \tag{12.215}$$

The integral of a measurable function is in general a number from $\overline{\mathbb{R}}$. A function $f\colon \mathrm{X} \longrightarrow \overline{\mathbb{R}}$ is called *integrable* or *summable* over X with respect to μ if it is measurable and $\int |f|\,d\mu < \infty$.

12.9.3.2 Some Properties of the Integral

Let $(\mathrm{X}, \mathcal{A}, \mu)$ be a measure space, $f, g : \mathrm{X} \longrightarrow \overline{\mathbb{R}}$ be measurable functions and $\alpha, \beta \in \mathbb{R}$.

1. If f is integrable, then f is finite a.e., i.e., $\mu\{x \in \mathrm{X}\colon |f(x)| = +\infty\} = 0$.

2. If f is integrable, then $\left|\int f\,d\mu\right| \leq \int |f|\,d\mu$.

3. If f is integrable and $f \geq 0$, then $\int f\,d\mu \geq 0$.

4. If $0 \leq g(x) \leq f(x)$ on X and f is integrable, then g is also integrable, and $\int g\,d\mu \leq \int f\,d\mu$.

5. If f, g are integrable, then $\alpha f + \beta g$ is integrable, and $\int (\alpha f + \beta g)\,d\mu = \alpha \int f\,d\mu + \beta \int g\,d\mu$.

6. If f, g are integrable on $A \in \mathcal{A}$, i.e., there exist the integrals $\int_A f\,d\mu$ and $\int_A g\,d\mu$ according to (12.215) and $f = g$ μ-a.e. on A, then $\int_A f\,d\mu = \int_A g\,d\mu$.

If $\mathrm{X} = \mathbb{R}^n$ and λ is the Lebesgue measure, then the introduced integral is the (n-dimensional) *Lebesgue integral* (see also 8.2.3.1, **3.**, p. 507). In the case $n = 1$ and $A = [a, b]$, for every continuous function

f on $[a,b]$ both the Riemann integral $\int_a^b f(x)\,dx$ (see 8.2.1.1, 2., p. 494) and the Lebesgue integral $\int_{[a,b]} f\,d\lambda$ are defined. Both values are finite and equal to each other. Furthermore, if f is a bounded Riemann integrable function on $[a,b]$, then it is also Lebesgue integrable and the values of the two integrals coincide.

The set of Lebesgue integrable functions is considerably larger than the set of the Riemann integrable functions and it has several advantages, e.g., when passing to the limit under the integral sign and f, $|f|$ are Lebesgue integrable simultaneously.

12.9.3.3 Convergence Theorems

Now Lebesgue measurable functions will be considered throughout.

1. B. Levi's Theorem on Monotone Convergence

Let $\{f_n\}_{n=1}^{\infty}$ be an a.e. monotone increasing sequence of non-negative integrable functions with values in $\overline{\mathbb{R}}$. Then

$$\lim_{n\to\infty}\int f_n\,d\mu = \int \lim_{n\to\infty} f_n\,d\mu. \tag{12.216}$$

2. Fatou's Theorem

Let $\{f_n\}_{n=1}^{\infty}$ be a sequence of non-negative $\overline{\mathbb{R}}$-valued measurable functions. Then

$$\int \liminf f_n\,d\mu \le \liminf \int f_n\,d\mu. \tag{12.217}$$

3. Lebesgue's Dominated Convergence Theorem

Let $\{f_n\}$ be a sequence of measurable functions convergent on X a.e. to some function f. If there exists an integrable function g such that $|f_n| \le g$ a.e., then $f = \lim\limits_{n\to\infty} f_n$ is integrable and there holds

$$\lim_{n\to\infty}\int f_n\,d\mu = \int \lim_{n\to\infty} f_n\,d\mu\,. \tag{12.218}$$

4. Radon-Nikodym Theorem

a) Assumptions: Let $(\mathrm{X}, \mathcal{A}, \mu)$ be a σ-finite measure space, i.e., there exists a sequence $\{A_n\}$, $A_n \in \mathcal{A}$ such that $\mathrm{X} = \bigcup\limits_{n=1}^{\infty} A_n$ and $\mu(A_n) < \infty$ for $\forall\, n$. In this case the measure is called σ *finite*. It is called *finite* if $\mu(\mathrm{X}) < \infty$, and it is called a *probability measure* if $\mu(\mathrm{X}) = 1$. A real function φ defined on $\mathcal{A}$ is called *absolutely continuous* with respect to μ if $\mu(A) = 0$ implies $\varphi(A) = 0$. This property is denoted by $\varphi \prec \mu$.
For an integrable function f, the function φ defined on $\mathcal{A}$ by $\varphi(A) = \int_A f\,d\mu$ is σ additive and absolutely continuous with respect to the measure μ. The converse of this property plays a fundamental role in many theoretical investigations and practical applications:

b) Radon–Nikodym Theorem: Suppose a σ-additive function φ and a measure μ are given on a σ algebra $\mathcal{A}$, and let $\varphi \prec \mu$. Then there exists a μ-integrable function f such that for each set $A \in \mathcal{A}$,

$$\varphi(A) = \int\limits_A f\,d\mu. \tag{12.219}$$

The function f is uniquely determined up to its equivalence class, and φ is non-negative if and only if $f \ge 0$ μ-a.e.

12.9.4 L^p Spaces

Let $(\mathrm{X}, \mathcal{A}, \mu)$ be a measure space and p a real number $1 \le p < \infty$. For a measurable function f, according to 12.9.2.2, p. 695, the function $|f|^p$ is measurable as well, so the expression

$$N_p(f) = \left(\int |f|^p \, d\mu \right)^{\frac{1}{p}} \tag{12.220}$$

is defined (and may be equal to $+\infty$). A measurable function $f\colon \mathrm{X} \longrightarrow \overline{\mathbb{R}}$ is called *p-th power integrable*, or an *L^p-function* if $N_p(f) < +\infty$ holds or, equivalent to this, if $|f|^p$ is integrable.

For every p with $1 \le p < +\infty$, the set of all L^p-functions, i.e., all functions p-th power integrable with respect to μ on X, is denoted by $\mathcal{L}^p(\mu)$ or by $\mathcal{L}^p(\mathrm{X})$ or in full detail $\mathcal{L}^p(\mathrm{X}, \mathcal{A}, \mu)$. For $p = 1$ the simple notation $\mathcal{L}(\mathrm{X})$ is used. For $p = 2$ the functions are called *quadratically integrable*.

The set of all measurable μ-a.e. bounded functions on X is denoted by $\mathcal{L}^\infty(\mu)$ and the essential supremum of a function f is defined as

$$N_\infty(f) = \text{ess. sup}\, f = \inf\{a \in \mathbb{R} : |f(x)| \le a \;\; \mu\text{–a.e.}\}. \tag{12.221}$$

$\mathcal{L}^p(\mu)$ $(1 \le p \le \infty)$ equipped with the usual operations for measurable functions and taking into consideration Minkowski inequality for integrals (see 1.4.2.13, p. 32), is a vector space and $N_p(\cdot)$ is a semi-norm on $\mathcal{L}^p(\mu)$. If $f \le g$ means that $f(x) \le g(x)$ holds μ-a.e., then $\mathcal{L}^p(\mu)$ is also a vector lattice and even a K-space (see 12.1.7.4, p. 660). Two functions $f, g \in \mathcal{L}^p(\mu)$ are called *equivalent* (or they are declared as equal) if $f = g$ μ-a.e. on X. In this way, functions are considered to be identical if they are equal μ-a.e. The factorization of the set $\mathcal{L}^p(\mathrm{X})$ modulo the linear subspace $N_p^{-1}(0)$ leads to a set of equivalence classes on which the algebraic operations and the order can be transferred naturally. So a vector lattice (K-space) is obtained again, which is denoted now by $L^p(\mathrm{X}, \mu)$ or $L^p(\mu)$. Its elements are called functions, as before, but actually they are classes of equivalent functions.

It is very important that $\|\hat{f}\|_p = N_p(f)$ is now a norm on $L^p(\mu)$ ($\hat{f}$ stands here for the equivalence class of f, which will later be denoted simply by f), and $(L^p(\mu), \|f\|_p)$ for every p with $1 \le p \le +\infty$ is a Banach lattice with several good compatibility conditions between norm and order. For $p = 2$ with $(f, g) = \displaystyle\int f\overline{g}\, d\mu$ as a scalar product, $L^2(\mu)$ is also a Hilbert space (see [12.12]).

Very often the space $L^p(\Omega)$ is considered for a measurable subset $\Omega \subset \mathbb{R}^n$. Its definition is not a problem because of step **5** in (12.9.3.1, p. 696).

The spaces $L^p(\Omega, \lambda)$, where λ is the n-dimensional Lebesgue measure, can also be introduced as the completions (see 12.2.2.5, p. 668 and 12.3.2, p. 670) of the non-complete normed spaces $\mathcal{C}(\Omega)$ of all continuous functions on the set $\Omega \subset \mathbb{R}^n$ equipped with the integral norm $\|x\|_p = \left(\int |x|^p \, d\lambda \right)^{\frac{1}{p}}$ $(1 \le p < \infty)$ (see [12.18]).

Let X be a set with a finite measure, i.e., $\mu(\mathrm{X}) < +\infty$, and suppose for the real numbers p_1, p_2, $1 \le p_1 < p_2 \le +\infty$. Then $L^{p_2}(\mathrm{X}, \mu) \subset L^{p_1}(\mathrm{X}, \mu)$, and with a constant $C = C(p_1, p_2, \mu(\mathrm{X})) > 0$ (independent of x), there holds the estimation $\|x\|_1 \le C\|x\|_2$ for $x \in L^{p_2}$ (here $\|x\|_k$ denotes the norm of the space $L^{p_k}(\mathrm{X}, \mu)$ $(k = 1, 2)$).

12.9.5 Distributions

12.9.5.1 Formula of Partial Integration

For an arbitrary (open) domain $\Omega \subseteq \mathbb{R}^n$, $\mathcal{C}_0^\infty(\Omega)$ denotes the set of all arbitrary many times in Ω differentiable functions φ with compact support, i.e., the set $\text{supp}(\varphi) = \overline{\{x \in \Omega : \varphi(x) \ne 0\}}$ is compact in $\mathbb{R}^n$ and lies in Ω. The set of all *locally summable* functions with respect to the Lebesgue measure in $\mathbb{R}^n$ is doted by $L^1_{loc}(\Omega)$, i.e., all the measurable functions f (equivalent classes) on Ω such that $\int_\omega |f|\, d\lambda < +\infty$ for every bounded domain $\omega \subset \Omega$.

Both sets are vector spaces (with the natural algebraic operations).

There hold $L^p(\Omega) \subset L^1_{loc}(\Omega)$ for $1 \le p \le \infty$, and $L^1_{loc}(\Omega) = L^1(\Omega)$ for a bounded Ω. If the elements of $\mathcal{C}^k(\overline{\Omega})$ are considered as the classes generated by them in $L^p(\Omega)$, then the inclusion $\mathcal{C}^k(\overline{\Omega}) \subset L^p(\Omega)$ holds for bounded Ω, where $\mathcal{C}^k(\overline{\Omega})$ is at once dense. If Ω is unbounded, then the set $\mathcal{C}_0^\infty(\Omega)$ is dense (in

this sense) in $L^p(\Omega)$.

For a given function $f \in \mathcal{C}^k(\overline{\Omega})$ and an arbitrary function $\varphi \in \mathcal{C}_0^\infty(\Omega)$ the formula of partial integration has the form

$$\int\limits_\Omega f(x)D^\alpha\varphi(x)\,d\lambda = (-1)^{|\alpha|}\int\limits_\Omega \varphi(x)D^\alpha f(x)\,d\lambda \tag{12.222}$$

$\forall\,\alpha$ with $|\alpha| \le k$ (the fact that $D^\alpha\varphi|_{\partial\Omega} = 0$ is used), and will be taken as the starting point for the definition of the generalized derivative of a function $f \in L^1_{loc}(\Omega)$.

12.9.5.2 Generalized Derivative

Suppose $f \in L^1_{loc}(\Omega)$. If there exists a function $g \in L^1_{loc}(\Omega)$ such that $\forall\,\varphi \in \mathcal{C}_0^\infty(\Omega)$ with respect to some multi-index α the equation

$$\int\limits_\Omega f(x)D^\alpha\varphi(x)\,d\lambda = (-1)^{|\alpha|}\int\limits_\Omega g(x)\varphi(x)\,d\lambda \tag{12.223}$$

holds, then g is called the *generalized derivative* (*derivative in the Sobolev sense* or *distributional derivative*) of order α of f. It is denoted by $g = D^\alpha f$ as in the classical case.

The convergence of a sequence $\{\varphi_k\}_{k=1}^\infty$ in the vector space $\mathcal{C}_0^\infty(\Omega)$ to $\varphi \in \mathcal{C}_0^\infty(\Omega)$ is defined as

$$\varphi_k \longrightarrow \varphi \ \text{ if and only if } \begin{cases} \text{a) } \exists \text{ a compact set } K \subset \Omega \text{ with } \operatorname{supp}(\varphi_k) \subset K \text{ for any } k\,, \\ \text{b) } D^\alpha\varphi_k \to D^\alpha\varphi \text{ uniformly on } K \text{ for each multi-index } \alpha\,. \end{cases} \tag{12.224}$$

The set $\mathcal{C}_0^\infty(\Omega)$, equipped with this convergence of sequences, is called the *fundamental space*, and is denoted by $\mathcal{D}(\Omega)$. Its elements are often called test functions.

12.9.5.3 Distributions

A linear functional ℓ on $\mathcal{D}(\Omega)$ continuous in the following sense (see 12.2.3, p. 668):

$$\varphi_k,\ \varphi \in \mathcal{D}(\Omega) \ \text{ and } \ \varphi_k \longrightarrow \varphi \ \text{ imply } \ \ell(\varphi_k) \longrightarrow \ell(\varphi) \tag{12.225}$$

is called a *generalized function* or a *distribution*.

■ **A:** If $f \in L^1_{loc}(\Omega)$, then

$$\ell_f(\varphi) = (f,\varphi) = \int_\Omega f(x)\varphi(x)\,d\lambda, \quad \varphi \in \mathcal{D}(\Omega) \tag{12.226}$$

is a distribution. A distribution, defined by a locally summable function as in (12.226), is called *regular*. Two regular distributions are equal, i.e., $\ell_f(\varphi) = \ell_g(\varphi)\ \forall\varphi \in \mathcal{D}(\Omega)$, if and only if $f = g$ a.e. with respect to λ.

■ **B:** Let $a \in \Omega$ be an arbitrary fixed point. Then $\ell_{\delta_a}(\varphi) = \varphi(a),\ \varphi \in \mathcal{D}(\Omega)$ is a linear continuous functional on $\mathcal{D}(\Omega)$, hence a distribution, which is called the Dirac distribution, δ distribution or δ function.

Since ℓ_{δ_a} cannot be generated by any locally summable function (see [12.11], [12.24]), it is an example for a non-regular distribution.

The set of all distributions is denoted by $\mathcal{D}'(\Omega)$. From a more general duality theory than that discussed in 12.5.4, p. 681, $\mathcal{D}'(\Omega)$ can be obtained as the dual space of $\mathcal{D}(\Omega)$. Consequently, one should write $\mathcal{D}^*(\Omega)$ instead. In the space $\mathcal{D}'(\Omega)$, it is possible to define several operations with its elements and with functions from $\mathcal{C}^\infty(\Omega)$, e.g., the derivative of a distribution or the convolution of two distributions, which make $\mathcal{D}'(\Omega)$ important not only in theoretical investigations but also in practical applications in electrical engineering, mechanics, etc.

For a review and for simple examples in applications of generalized functions see, e.g., [12.11],[12.24].

Here, only the notion of the derivative of a generalized function is discussed.

12.9.5.4 Derivative of a Distribution

If ℓ is a given distribution, then the distribution $D^\alpha \ell$ defined by

$$(D^\alpha \ell)(\varphi) = (-1)^{|\alpha|}\ell(D^\alpha \varphi), \quad \varphi \in D(\Omega), \tag{12.227}$$

is called the *distributional derivative* of order α of ℓ.
Let f be a continuously differentiable function, say on $\mathbb{R}$ (so f is locally summable on $\mathbb{R}$, and f can be considered as a distribution), let f' be its classical derivative and $D^1 f$ its distributional derivative of order 1. Then:

$$(D^1 f, \varphi) = \int_{\mathrm{R}} f'(x)\varphi(x)\,dx, \tag{12.228a}$$

from which by partial integration there follows

$$(D^1 f, \varphi) = -\int_{\mathrm{R}} f(x)\varphi'(x)\,dx = -(f, \varphi'). \tag{12.228b}$$

In the case of a regular distribution ℓ_f with $f \in L^1_{loc}(\Omega)$ by using (12.226)

$$(D^\alpha \ell_f)(\varphi) = (-1)^{|\alpha|}\ell_f(D^\alpha \varphi) = (-1)^{|\alpha|}\int_\Omega f(x) D^\alpha \varphi\, d\lambda \tag{12.229}$$

is obtained, which is the generalized derivative of the function f in the Sobolev sense (see (12.223)).

■ **A:** For the regular distribution generated by the (obviously locally summable) Heaviside function

$$\Theta(x) = \begin{cases} 1 \text{ for } x \geq 0, \\ 0 \text{ for } x < 0 \end{cases} \tag{12.230}$$

the non-regular δ distribution is obtained as the derivative.

■ **B:** In mathematical modeling of technical and physical problems one is faced with (in a certain sense idealized) influences concentrated at one point, such as a “point-like” force, needle-deflection, collision, etc., which can be expressed mathematically by using the δ or Heaviside function. For example, $m\delta_a$ is the mass density of a point-like mass m concentrated at one point a $(0 \leq a \leq l)$ of a beam of length l.
The motion of a spring-mass system on which at time t_0 there acts a momentary external force F is described by the equation $\ddot{x} + \omega^2 x = F\delta_{t_0}$. With the initial conditions $x(0) = \dot{x}(0) = 0$ its solution is $x(t) = \frac{F}{\omega}\sin(\omega(t - t_0))\Theta(t - t_0)$.

13 Vector Analysis and Vector Fields

13.1 Basic Notions of the Theory of Vector Fields

13.1.1 Vector Functions of a Scalar Variable

13.1.1.1 Definitions

1. Vector Function of a Scalar Variable t

A vector function of a scalar variable is a vector $\vec{\mathbf{a}}$ whose components are real functions of t:

$$\vec{\mathbf{a}} = \vec{\mathbf{a}}(t) = a_x(t)\vec{\mathbf{e}}_x + a_y(t)\vec{\mathbf{e}}_y + a_z(t)\vec{\mathbf{e}}_z. \tag{13.1}$$

The notions of limit, continuity, differentiability are defined componentwise for the vector $\vec{\mathbf{a}}(t)$.

2. Hodograph of a Vector Function

Considering the vector function $\vec{\mathbf{a}}(t)$ as a position or radius vector $\vec{\mathbf{r}} = \vec{\mathbf{r}}(t)$ of a point P, then this function describes a space curve while t varies **(Fig. 13.1)**. This space curve is called the *hodograph* of the vector function $\vec{\mathbf{a}}(t)$.

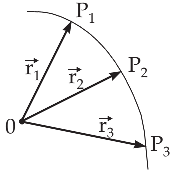


Figure 13.1

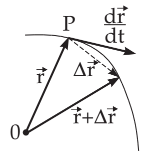


Figure 13.2

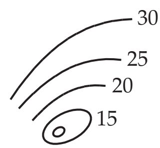


Figure 13.3

13.1.1.2 Derivative of a Vector Function

The derivative of (13.1) with respect to t is also a vector function of t:

$$\frac{d\vec{\mathbf{a}}}{dt} = \lim_{\Delta t \to 0} \frac{\vec{\mathbf{a}}(t+\Delta t) - \vec{\mathbf{a}}(t)}{\Delta t} = \frac{da_x(t)}{dt}\vec{\mathbf{e}}_x + \frac{da_y(t)}{dt}\vec{\mathbf{e}}_y + \frac{da_z(t)}{dt}\vec{\mathbf{e}}_z. \tag{13.2}$$

The geometric representation of the derivative $\frac{d\vec{\mathbf{r}}}{dt}$ of the radius vector is a vector pointing in the direction of the tangent of the hodograph at the point P **(Fig. 13.2)**. Its length depends on the choice of the parameter t. If t is the time, then the vector $\vec{\mathbf{r}}(t)$ describes the motion of a point P in space (the space curve is its path), and $\frac{d\vec{\mathbf{r}}}{dt}$ has the direction and magnitude of the velocity of this motion. If $t = s$ is the arclength of this space curve, measured from a certain point, then obviously $\left|\frac{d\vec{\mathbf{r}}}{ds}\right| = 1$.

13.1.1.3 Rules of Differentiation for Vectors

$$\frac{d}{dt}\,(\vec{\mathbf{a}} \pm \vec{\mathbf{b}} \pm \vec{\mathbf{c}}) = \frac{d\vec{\mathbf{a}}}{dt} \pm \frac{d\vec{\mathbf{b}}}{dt} \pm \frac{d\vec{\mathbf{c}}}{dt}, \tag{13.3a}$$

$$\frac{d}{dt}\,(\varphi\vec{\mathbf{a}}) = \frac{d\varphi}{dt}\vec{\mathbf{a}} + \varphi\frac{d\vec{\mathbf{a}}}{dt} \qquad (\varphi \text{ is a scalar function of } t), \tag{13.3b}$$

$$\frac{d}{dt}\,(\vec{\mathbf{a}}\vec{\mathbf{b}}) = \frac{d\vec{\mathbf{a}}}{dt}\vec{\mathbf{b}} + \vec{\mathbf{a}}\frac{d\vec{\mathbf{b}}}{dt}, \tag{13.3c}$$

I.N. Bronshtein et al., *Handbook of Mathematics*,
DOI 10.1007/978-3-662-46221-8_13

$$\frac{d}{dt}(\vec{\mathbf{a}}\times\vec{\mathbf{b}}) = \frac{d\vec{\mathbf{a}}}{dt}\times\vec{\mathbf{b}} + \vec{\mathbf{a}}\times\frac{d\vec{\mathbf{b}}}{dt} \quad \text{(the factors must not be interchanged)}, \tag{13.3d}$$

$$\frac{d}{dt}\vec{\mathbf{a}}[\varphi(t)] = \frac{d\vec{\mathbf{a}}}{d\varphi}\cdot\frac{d\varphi}{dt} \quad \text{(chain rule)}. \tag{13.3e}$$

If $|\vec{\mathbf{a}}(t)| = \text{const}$, i.e., $\vec{\mathbf{a}}^2(t) = \vec{\mathbf{a}}(t)\cdot\vec{\mathbf{a}}(t) = \text{const}$, then it follows from (13.3c) that $\vec{\mathbf{a}}\cdot\dfrac{d\vec{\mathbf{a}}}{dt} = 0$, i.e., $\dfrac{d\vec{\mathbf{a}}}{dt}$ and $\vec{\mathbf{a}}$ are perpendicular to each other. Examples of this fact:

■ **A:** Radius and tangent vectors of a circle in the plane and

■ **B:** position and tangent vectors of a curve on the sphere. Then the hodograph is a *spherical curve*.

13.1.1.4 Taylor Expansion for Vector Functions

$$\vec{\mathbf{a}}(t+h) = \vec{\mathbf{a}}(t) + h\frac{d\vec{\mathbf{a}}}{dt} + \frac{h^2}{2!}\frac{d^2\vec{\mathbf{a}}}{dt^2} + \cdots + \frac{h^n}{n!}\frac{d^n\vec{\mathbf{a}}}{dt^n} + \cdots. \tag{13.4}$$

The expansion of a vector function in a Taylor series makes sense only if it is convergent. Because the limit is defined componentwise, the convergence can be checked componentwise, so the convergence of this series with vector terms can be determined exactly by the same methods as the convergence of a series with complex terms (see 14.3.2, p. 751). So the convergence of a series with vector terms is reduced to the convergence of a series with scalar terms.
The differential of a vector function $\vec{\mathbf{a}}(t)$ is defined by:

$$d\vec{\mathbf{a}} = \frac{d\vec{\mathbf{a}}}{dt}\Delta t. \tag{13.5}$$

13.1.2 Scalar Fields

13.1.2.1 Scalar Field or Scalar Point Function

If a number (scalar value) U is assigned to every point P of a subset of space, then one writes

$$U = U(P) \tag{13.6a}$$

and one calls (13.6a) a *scalar field* (or *scalar function*).

■ Examples of scalar fields are temperature, density, potential, etc., of solids.

A scalar field $U = U(P)$ can also be considered as

$$U = U(\vec{\mathbf{r}}), \tag{13.6b}$$

where $\vec{\mathbf{r}}$ is the position vector of the point P with a given pole 0 (see 3.5.1.1, **6.**, p. 182).

13.1.2.2 Important Special Cases of Scalar Fields

1. Plane Field

One speaks of a plane field, if the function is defined only for the points of a plane in space.

2. Central Field

If a function has the same value at all points P lying at the same distance from a fixed point $C(\vec{\mathbf{r}}_1)$, called the center, then it is called a *central symmetric field* or also a *central or spherical field*. The function U depends only on the distance $\overline{CP} = |\vec{\mathbf{r}}|$:

$$U = f(|\vec{\mathbf{r}}|). \tag{13.7a}$$

■ The field of the intensity of a point-like source, e.g., the field of brightness of a point-like source of light at the pole, can be described with $|\vec{\mathbf{r}}| = r$ as the distance from the light source:

$$U = \frac{c}{r^2} \quad (c\ \text{const}). \tag{13.7b}$$

3. Axial Field

If the function U has the same value at all points lying at an equal distance from a certain straight line (axis of the field) then the field is called *cylindrically symmetric* or an *axially symmetric field*, or briefly an *axial field*.

13.1.2.3 Coordinate Representation of Scalar Fields

If the points of a subset of space are given by their coordinates, e.g., by Cartesian, cylindrical, or spherical coordinates, then the corresponding scalar field (13.6a) is represented, in general, by a function of three variables:

$$U = \Phi(x, y, z), \qquad U = \Psi(\rho, \varphi, z) \qquad \text{or} \qquad U = \chi(r, \vartheta, \varphi). \tag{13.8a}$$

In the case of a plane field, a function with two variables is sufficient. It has the form in Cartesian and polar coordinates:

$$U = \Phi(x, y) \qquad \text{or} \qquad U = \Psi(\rho, \varphi). \tag{13.8b}$$

The functions in (13.8a) and (13.8b), in general, are assumed to be continuous, except, maybe, at some points, curves or surfaces of discontinuity. The functions have the form

a) for a central field: $$U = U(\sqrt{x^2 + y^2 + z^2}) = U(\sqrt{\rho^2 + z^2}) = U(r) \tag{13.9a}$$

with the origin of the coordinate system as the *pole* of the field,

b) for an axial field: $$U = U(\sqrt{x^2 + y^2}) = U(\rho) = U(r \sin \vartheta) \tag{13.9b}$$

with the z-axis as the axis of the field.
Dealing with central fields is easiest using spherical coordinates, with axial fields using cylindrical coordinates.

13.1.2.4 Level Surfaces and Level Lines of a Field

1. Level Surface

A level surface is the union of all points P in space where the function (13.6a) has a constant value

$$U = U(P) = \text{const.} \tag{13.10a}$$

Different constants $U_0, U_1, U_2, \ldots$ define different level surfaces. There is a level surface passing through every point except the points where the function is not defined. The level surface equations in the three coordinate systems used so far are:

$$U = \Phi(x, y, z) = \text{const}, \qquad U = \Psi(\rho, \varphi, z) = \text{const}, \qquad U = \chi(r, \vartheta, \varphi) = \text{const}. \tag{13.10b}$$

■ Examples of level surfaces of different fields:

A: $U = \vec{\mathbf{c}}\,\vec{\mathbf{r}} = c_x x + c_y y + c_z z$: Parallel planes.
B: $U = x^2 + 2y^2 + 4z^2$: Similar ellipsoids in similar positions.
C: Central field: Concentric spheres.
D: Axial field: Coaxial cylinders.

2. Level Lines

Level lines replace level surfaces in plane fields. They satisfy the equation

$$U = \text{const.} \tag{13.11}$$

Level lines are usually drawn for equal intervals of U and each of them is marked by the corresponding value of U **(Fig. 13.3)**.

■ Well-known examples are the isobaric lines on a synoptic map or the contour lines on topographic maps.

In particular cases, level surfaces degenerate into points or lines, and level lines degenerate into separate points.

■ The level lines of the fields a) $U = xy$, b) $U = \dfrac{y}{x^2}$, c) $U = x^2 + y^2 = \rho^2$, d) $U = \dfrac{1}{\rho}$ are represented in **Fig. 13.4**.

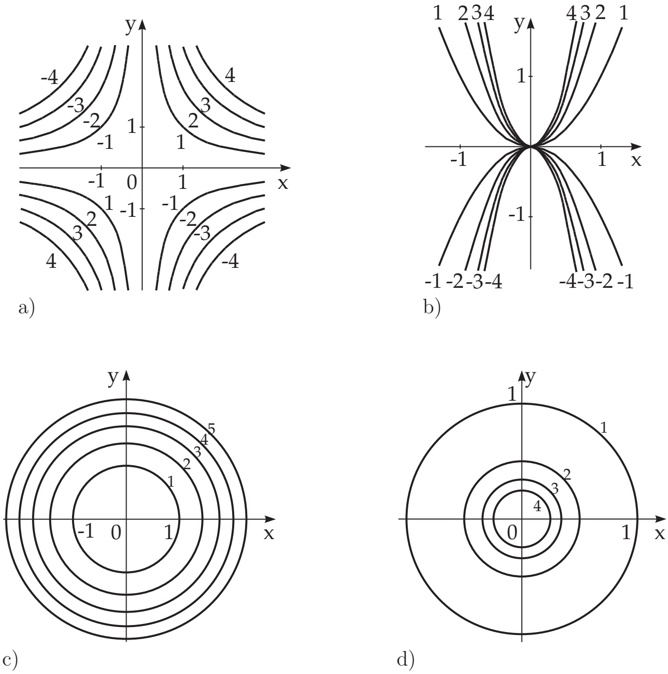


Figure 13.4

13.1.3 Vector Fields

13.1.3.1 Vector Field or Vector Point Function

If a vector $\vec{\mathbf{V}}$ is assigned to every point P of a subset of space, then it is denoted by

$$\vec{\mathbf{V}} = \vec{\mathbf{V}}(P) \tag{13.12a}$$

and calls (13.12a) a *vector field*.

■ Examples of vector fields are the velocity field of a fluid in motion, a field of force, and a magnetic or electric intensity field.

A vector field $\vec{\mathbf{V}} = \vec{\mathbf{V}}(P)$ can be regarded as a vector function

$$\vec{\mathbf{V}} = \vec{\mathbf{V}}(\vec{\mathbf{r}}), \tag{13.12b}$$

where $\vec{\mathbf{r}}$ is the position vector of the point P with a given pole 0. If all values of $\vec{\mathbf{r}}$ as well as $\vec{\mathbf{V}}$ lie in a plane, then the field is called a plane vector field (see 3.5.2, p. 190).

13.1.3.2 Important Cases of Vector Fields

1. Central Vector Field

In a central vector field all vectors $\vec{\mathbf{V}}$ lie on straight lines passing through a fixed point called the *center* **(Fig. 13.5a)**.
Locating the pole at the center, then the field is defined by the formula

$$\vec{\mathbf{V}} = f(\vec{\mathbf{r}})\,\vec{\mathbf{r}}, \tag{13.13a}$$

where all the vectors have the same direction as the radius vector $\vec{\mathbf{r}}$. It often has some advantage to define the field by the formula

$$\vec{\mathbf{V}} = \varphi(\vec{\mathbf{r}})\frac{\vec{\mathbf{r}}}{r} \quad (r = |\vec{\mathbf{r}}|), \tag{13.13b}$$

where $|\varphi(\vec{\mathbf{r}})|$ is the length of the vector $\vec{\mathbf{V}}$ and $\frac{\vec{\mathbf{r}}}{r}$ is a unit vector into the direction of $\vec{\mathbf{r}}$.

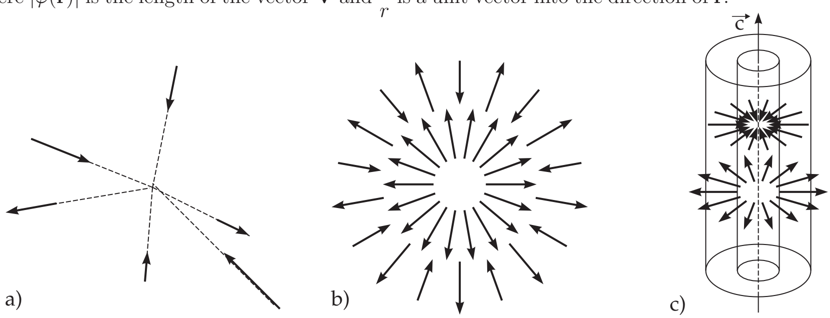


Figure 13.5

2. Spherical Vector Field

A spherical vector field is a special case of a central vector field, where the length of the vector $\vec{\mathbf{V}}$ depends only on the distance $|\vec{\mathbf{r}}|$ **(Fig. 13.5b)**.

■ Examples are the *Newton* and the *Coulomb force field* of a point-like mass or of a point-like electric charge:

$$\vec{\mathbf{V}} = \frac{c}{r^3}\vec{\mathbf{r}} = \frac{c}{r^2}\frac{\vec{\mathbf{r}}}{r} \quad (c \text{ const}). \tag{13.14}$$

The special case of a plane spherical vector field is called a *circular field*.

3. Cylindrical Vector Field

a) All vectors $\vec{\mathbf{V}}$ lie on straight lines intersecting a certain line (called the *axis*) and perpendicular to it, and

b) all vectors $\vec{\mathbf{V}}$ at the points lying at the same distance from the axis have equal length, and they are directed either toward the axis or away from it **(Fig. 13.5c)**.

Locating the pole on the axis parallel to the unit vector $\vec{\mathbf{c}}$, then the field has the form

$$\vec{\mathbf{V}} = \varphi(\rho)\frac{\vec{\mathbf{r}}^*}{\rho}, \tag{13.15a}$$

where $\vec{\mathbf{r}}^*$ is the projection of $\vec{\mathbf{r}}$ on a plane perpendicular to the axis:

$$\vec{\mathbf{r}}^* = \vec{\mathbf{c}} \times (\vec{\mathbf{r}} \times \vec{\mathbf{c}}). \tag{13.15b}$$

By intersecting this field with planes perpendicular to the axis, one always gets equal circular fields.

13.1.3.3 Coordinate Representation of Vector Fields

1. Vector Field in Cartesian Coordinates

The vector field (13.12a) can be defined by scalar fields $V_1(\vec{\mathbf{r}})$, $V_2(\vec{\mathbf{r}})$, and $V_3(\vec{\mathbf{r}})$ which are the coordinate functions of $\vec{\mathbf{V}}$, i.e., the coefficients of its decomposition into any three non-coplanar base vectors $\vec{\mathbf{e}}_1$, $\vec{\mathbf{e}}_2$, and $\vec{\mathbf{e}}_3$:

$$\vec{\mathbf{V}} = V_1\vec{\mathbf{e}}_1 + V_2\vec{\mathbf{e}}_2 + V_3\vec{\mathbf{e}}_3. \tag{13.16a}$$

With the coordinate unit vectors $\vec{\mathbf{i}}, \vec{\mathbf{j}}, \vec{\mathbf{k}}$ as base vectors and expressing the coefficients V_1, V_2, V_3 in Cartesian coordinates one gets

$$\vec{\mathbf{V}} = V_x(x,y,z)\vec{\mathbf{i}} + V_y(x,y,z)\vec{\mathbf{j}} + V_z(x,y,z)\vec{\mathbf{k}}. \tag{13.16b}$$

So, the vector field can be defined with the help of three scalar functions of three scalar variables.

2. Vector Field in Cylindrical and Spherical Coordinates

In cylindrical and spherical coordinates, the coordinate unit vectors

$$\vec{\mathbf{e}}_\rho,\ \vec{\mathbf{e}}_\varphi,\ \vec{\mathbf{e}}_z\ (= \vec{\mathbf{k}}), \qquad \text{and} \qquad \vec{\mathbf{e}}_r\ (= \frac{\vec{\mathbf{r}}}{r}),\ \vec{\mathbf{e}}_\vartheta,\ \vec{\mathbf{e}}_\varphi \tag{13.17a}$$

are tangents to the coordinate lines at each point **(Fig. 13.6, 13.7)**. In this order they always form a right-handed system. The coefficients are expressed as functions of the corresponding coordinates:

$$\vec{\mathbf{V}} = V_\rho(\rho,\varphi,z)\vec{\mathbf{e}}_\rho + V_\varphi(\rho,\varphi,z)\vec{\mathbf{e}}_\varphi + V_z(\rho,\varphi,z)\vec{\mathbf{e}}_z\,, \tag{13.17b}$$

$$\vec{\mathbf{V}} = V_r(r,\vartheta,\varphi)\vec{\mathbf{e}}_r + V_\varphi(r,\vartheta,\varphi)\vec{\mathbf{e}}_\varphi + V_\vartheta(r,\vartheta,\varphi)\vec{\mathbf{e}}_\vartheta. \tag{13.17c}$$

At transition from one point to the other, the coordinate unit vectors change their directions, but remain mutually perpendicular.

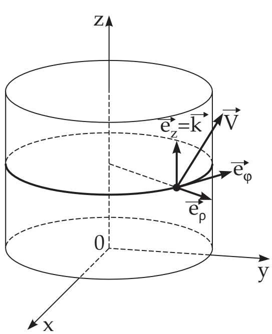


Figure 13.6

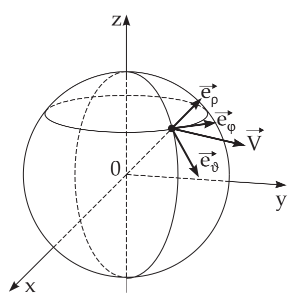


Figure 13.7

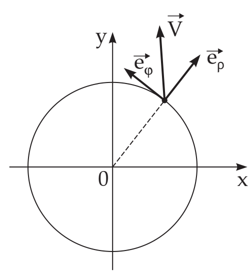


Figure 13.8

13.1.3.4 Transformation of Coordinate Systems

See also **Table 13.1**.

1. Cartesian Coordinates in Terms of Cylindrical Coordinates

$$V_x = V_\rho\cos\varphi - V_\varphi\sin\varphi, \qquad V_y = V_\rho\sin\varphi + V_\varphi\cos\varphi, \qquad V_z = V_z. \tag{13.18}$$

2. Cylindrical Coordinates in Terms of Cartesian Coordinates

$$V_\rho = V_x \cos\varphi + V_y \sin\varphi, \qquad V_\varphi = -V_x \sin\varphi + V_y \cos\varphi, \qquad V_z = V_z. \tag{13.19}$$

3. Cartesian Coordinates in Terms of Spherical Coordinates

$$\begin{aligned} V_x &= V_r \sin\vartheta \cos\varphi - V_\varphi \sin\varphi + V_\vartheta \cos\varphi \cos\vartheta, \\ V_y &= V_r \sin\vartheta \sin\varphi + V_\varphi \cos\varphi + V_\vartheta \sin\varphi \cos\vartheta, \\ V_z &= V_r \cos\vartheta - V_\vartheta \sin\vartheta. \end{aligned} \tag{13.20}$$

4. Spherical Coordinates in Terms of Cartesian Coordinates

$$\begin{aligned} V_r &= V_x \sin\vartheta \cos\varphi + V_y \sin\vartheta \sin\varphi + V_z \cos\vartheta, \\ V_\vartheta &= V_x \cos\vartheta \cos\varphi + V_y \cos\vartheta \sin\varphi - V_z \sin\vartheta, \\ V_\varphi &= -V_x \sin\varphi + V_y \cos\varphi. \end{aligned} \tag{13.21}$$

5. Expression of a Spherical Vector Field in Cartesian Coordinates

$$\vec{\mathbf{V}} = \varphi(\sqrt{x^2+y^2+z^2})(x\vec{\mathbf{i}} + y\vec{\mathbf{j}} + z\vec{\mathbf{k}}). \tag{13.22}$$

6. Expression of a Cylindrical Vector Field in Cartesian Coordinates

$$\vec{\mathbf{V}} = \varphi(\sqrt{x^2+y^2})(x\vec{\mathbf{i}} + y\vec{\mathbf{j}}). \tag{13.23}$$

In the case of a spherical vector field, spherical coordinates are most convenient for investigations, i.e., the form $\vec{\mathbf{V}} = V(r)\vec{\mathbf{e}}_r$; and for investigations in cylindrical fields, cylindrical coordinates are most convenient, i.e., the form $\vec{\mathbf{V}} = V(\varphi)\vec{\mathbf{e}}_\varphi$. In the case of a plane field **(Fig. 13.8)**

$$\vec{\mathbf{V}} = V_x(x,y)\vec{\mathbf{i}} + V_y(x,y)\vec{\mathbf{j}} = V_\rho(\rho,\varphi)\vec{\mathbf{e}}_\rho + V_\varphi(\rho,\varphi)\vec{\mathbf{e}}_\varphi, \tag{13.24}$$

holds and for a circular field

$$\vec{\mathbf{V}} = \varphi(\sqrt{x^2+y^2})(x\vec{\mathbf{i}} + y\vec{\mathbf{j}}) = \varphi(\rho)\vec{\mathbf{e}}_\rho. \tag{13.25}$$

Table 13.1 Relations between the components of a vector in Cartesian, cylindrical, and spherical coordinates

Cartesian coordinates	**Cylindrical coord.**	**Spherical coordinates**
$\vec{\mathbf{V}} = V_x\vec{\mathbf{e}}_x + V_y\vec{\mathbf{e}}_y + V_z\vec{\mathbf{e}}_z$	$V_\rho\vec{\mathbf{e}}_\rho + V_\varphi\vec{\mathbf{e}}_\varphi + V_z\vec{\mathbf{e}}_z$	$V_r\vec{\mathbf{e}}_r + V_\vartheta\vec{\mathbf{e}}_\vartheta + V_\varphi\vec{\mathbf{e}}_\varphi$
V_x	$= V_\rho \cos\varphi - V_\varphi \sin\varphi$	$= V_r \sin\vartheta \cos\varphi + V_\vartheta \cos\vartheta \cos\varphi - V_\varphi \sin\varphi$
V_y	$= V_\rho \sin\varphi + V_\varphi \cos\varphi$	$= V_r \sin\vartheta \sin\varphi + V_\vartheta \cos\vartheta \sin\varphi + V_\varphi \cos\varphi$
V_z	$= V_z$	$= V_r \cos\vartheta - V_\vartheta \sin\vartheta$
$V_x \cos\varphi + V_y \sin\varphi$	$= V_\rho$	$= V_r \sin\vartheta + V_\vartheta \cos\vartheta$
$-V_x \sin\varphi + V_y \cos\varphi$	$= V_\varphi$	$= V_\varphi$
V_z	$= V_z$	$= V_r \cos\vartheta - V_\vartheta \sin\vartheta$
$V_x \sin\vartheta \cos\varphi + V_y \sin\vartheta \sin\varphi + V_z \cos\vartheta$	$= V_\rho \sin\vartheta + V_z \cos\vartheta$	$= V_r$
$V_x \cos\vartheta \cos\varphi + V_y \cos\vartheta \sin\varphi - V_z \sin\varphi$	$= V_\rho \cos\vartheta - V_z \sin\vartheta$	$= V_\vartheta$
$-V_x \sin\varphi + V_y \cos\varphi$	$= V_\varphi$	$= V_\varphi$

13.1.3.5 Vector Lines

A curve C is called a *line of a vector* or a *vector line* of the vector field $\vec{\mathbf{V}}(\vec{\mathbf{r}})$ **(Fig. 13.9)** if the vector $\vec{\mathbf{V}}(\vec{\mathbf{r}})$ is a tangent vector of the curve at every point P. There is a vector line passing through every point of the field. Vector lines do not intersect each other, except, maybe, at points where the function $\vec{\mathbf{V}}$ is not defined, or where it is the zero vector. The differential equations of the vector lines of a vector field $\vec{\mathbf{V}}$ given in Cartesian coordinates are

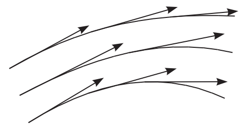


Figure 13.9

a) in general: $\dfrac{dx}{V_x} = \dfrac{dy}{V_y} = \dfrac{dz}{V_z}$, (13.26a) **b) for a plane field:** $\dfrac{dx}{V_x} = \dfrac{dy}{V_y}$. (13.26b)

To solve these differential equations see 9.1.1.2, p. 542 or 9.2.1.1, p. 571.

■ **A:** The vector lines of a central field are rays starting at the center of the vector field.

■ **B:** The vector lines of the vector field $\vec{\mathbf{V}} = \vec{\mathbf{c}} \times \vec{\mathbf{r}}$ are circles lying in planes perpendicular to the vector $\vec{\mathbf{c}}$. Their centers are on the axis parallel to $\vec{\mathbf{c}}$.

13.2 Differential Operators of Space

13.2.1 Directional and Space Derivatives

13.2.1.1 Directional Derivative of a Scalar Field

The directional derivative of a scalar field $U = U(\vec{\mathbf{r}})$ at a point P with position vector $\vec{\mathbf{r}}$ in the direction $\vec{\mathbf{c}}$ **(Fig. 13.10)** is defined as the limit of the quotient

$$\frac{\partial U}{\partial \vec{\mathbf{c}}} = \lim_{\varepsilon \to 0} \frac{U(\vec{\mathbf{r}} + \varepsilon\vec{\mathbf{c}}) - U(\vec{\mathbf{r}})}{\varepsilon}. \tag{13.27}$$

If the derivative of the field $U = U(\vec{\mathbf{r}})$ at a point $\vec{\mathbf{r}}$ in the direction of the unit vector $\vec{\mathbf{c}}^0$ of $\vec{\mathbf{c}}$ is denoted by $\dfrac{\partial U}{\partial \vec{\mathbf{c}}^0}$, then the relation between the derivative of the function with respect to the vector $\vec{\mathbf{c}}$ and with respect to its unit vector $\vec{\mathbf{c}}^0$ at the same point is

$$\frac{\partial U}{\partial \vec{\mathbf{c}}} = |\vec{\mathbf{c}}| \frac{\partial U}{\partial \vec{\mathbf{c}}^0}. \tag{13.28}$$

The derivative $\dfrac{\partial U}{\partial \vec{\mathbf{c}}^0}$ with respect to the unit vector represents the speed of increase of the function U in the direction of the vector $\vec{\mathbf{c}}^0$ at the point $\vec{\mathbf{r}}$. If $\vec{\mathbf{n}}$ is the normal unit vector to the level surface passing through the point $\vec{\mathbf{r}}$, and $\vec{\mathbf{n}}$ is pointing in the direction of increasing U, then $\dfrac{\partial U}{\partial \vec{\mathbf{n}}}$ has the greatest value among all the derivatives at the point with respect to the unit vectors in different directions. Between the directional derivatives with respect to $\vec{\mathbf{n}}$ and with respect to any direction $\vec{\mathbf{c}}^0$ holds the relation

$$\frac{\partial U}{\partial \vec{\mathbf{c}}^0} = \frac{\partial U}{\partial \vec{\mathbf{n}}} \cos(\vec{\mathbf{c}}^0, \vec{\mathbf{n}}) = \frac{\partial U}{\partial \vec{\mathbf{n}}} \cos\varphi = \vec{\mathbf{c}}^0 \cdot \operatorname{grad} U \quad \text{(see (13.34), p. 710)}. \tag{13.29}$$

Hereafter, directional derivatives always mean the directional derivative with respect to a unit vector.

13.2.1.2 Directional Derivative of a Vector Field

The directional derivative of a vector field is defined analogously to the directional derivative of a scalar field. The directional derivative of the vector field $\vec{\mathbf{V}} = \vec{\mathbf{V}}(\vec{\mathbf{r}})$ at a point P with position vector $\vec{\mathbf{r}}$

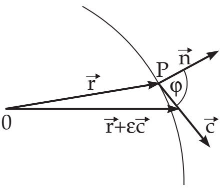


Figure 13.10

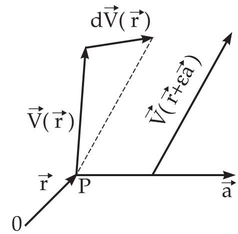


Figure 13.11

(Fig. 13.11) with respect to the vector $\vec{\mathbf{a}}$ is defined as the limit of the quotient

$$\frac{\partial \vec{\mathbf{V}}}{\partial \vec{\mathbf{a}}} = \lim_{\varepsilon \to 0} \frac{\vec{\mathbf{V}}(\vec{\mathbf{r}} + \varepsilon \vec{\mathbf{a}}) - \vec{\mathbf{V}}(\vec{\mathbf{r}})}{\varepsilon} . \tag{13.30}$$

If the derivative of the vector field $\vec{\mathbf{V}} = \vec{\mathbf{V}}(\vec{\mathbf{r}})$ at a point $\vec{\mathbf{r}}$ in the direction of the unit vector $\vec{\mathbf{a}}^0$ of $\vec{\mathbf{a}}$ is denoted by $\frac{\partial \vec{\mathbf{V}}}{\partial \vec{\mathbf{a}}^0}$, then

$$\frac{\partial \vec{\mathbf{V}}}{\partial \vec{\mathbf{a}}} = |\vec{\mathbf{a}}| \frac{\partial \vec{\mathbf{V}}}{\partial \vec{\mathbf{a}}^0} . \tag{13.31}$$

In Cartesian coordinates, i.e., for $\vec{\mathbf{V}} = V_x \vec{\mathbf{e}}_x + V_y \vec{\mathbf{e}}_y + V_z \vec{\mathbf{e}}_z$, $\vec{\mathbf{a}} = a_x \vec{\mathbf{e}}_x + a_y \vec{\mathbf{e}}_y + a_z \vec{\mathbf{e}}_z$, holds:

$$(\vec{\mathbf{a}} \cdot \operatorname{grad}) \vec{\mathbf{V}} = (\vec{\mathbf{a}} \cdot \operatorname{grad} V_x) \vec{\mathbf{e}_x} + (\vec{\mathbf{a}} \cdot \operatorname{grad} V_y) \vec{\mathbf{e}_y} + (\vec{\mathbf{a}} \cdot \operatorname{grad} V_z) \vec{\mathbf{e}_z} , \tag{13.32a}$$

in general coordinates:

$$(\vec{\mathbf{a}} \cdot \operatorname{grad}) \vec{\mathbf{V}} = \frac{1}{2} (\operatorname{rot} (\vec{\mathbf{V}} \times \vec{\mathbf{a}}) + \operatorname{grad} (\vec{\mathbf{a}} \cdot \vec{\mathbf{V}}) + \vec{\mathbf{a}} \operatorname{div} \vec{\mathbf{V}} - \vec{\mathbf{V}} \operatorname{div} \vec{\mathbf{a}} - \vec{\mathbf{a}} \times \operatorname{rot} \vec{\mathbf{V}} - \vec{\mathbf{V}} \times \operatorname{rot} \vec{\mathbf{a}} . \tag{13.32b}$$

13.2.1.3 Volume Derivative

Volume derivatives of a scalar field $U = U(\vec{\mathbf{r}})$ or a vector field $\vec{\mathbf{V}}$ at a point $\vec{\mathbf{r}}$ are quantities of three forms, which are obtained as follows:

1. Surrounding the point $\vec{\mathbf{r}}$ of the scalar field or of the vector field by a closed surface Σ. This surface can be represented in parametric form $\vec{\mathbf{r}} = \vec{\mathbf{r}}(u, v) = x(u, v) \vec{\mathbf{e}}_x + y(u, v) \vec{\mathbf{e}}_y + z(u, v) \vec{\mathbf{e}}_z$, so the corresponding vectorial surface element is

$$d\vec{\mathbf{S}} = \frac{\partial \vec{\mathbf{r}}}{\partial u} \times \frac{\partial \vec{\mathbf{r}}}{\partial v} \, du \, dv . \tag{13.33a}$$

2. Evaluating the surface integral over the closed surface Σ . Here, the following three types of integrals can be considered:

$$\oiint_{(\Sigma)} U \, d\vec{\mathbf{S}} , \qquad \oiint_{(\Sigma)} \vec{\mathbf{V}} \cdot d\vec{\mathbf{S}} , \qquad \oiint_{(\Sigma)} \vec{\mathbf{V}} \times d\vec{\mathbf{S}} . \tag{13.33b}$$

3. Determining the limits (if they exist)

$$\lim_{V \to 0} \frac{1}{V} \oiint_{(\Sigma)} U \, d\vec{\mathbf{S}} , \quad \lim_{V \to 0} \frac{1}{V} \oiint_{(\Sigma)} \vec{\mathbf{V}} \cdot d\vec{\mathbf{S}} , \quad \lim_{V \to 0} \frac{1}{V} \oiint_{(\Sigma)} \vec{\mathbf{V}} \times d\vec{\mathbf{S}} . \tag{13.33c}$$

Here V denotes the volume of the region of space that contains the point with the position vector $\vec{\mathbf{r}}$ inside, and which is bounded by the considered closed surface Σ .

The limits (13.33c) are called volume derivatives. The *gradient of a scalar field* and the *divergence* and the *rotation* of a vector field can be derived from them in the given order. In the following paragraphs, these notions will be discussed in details (even defining them again.)

13.2.2 Gradient of a Scalar Field

The gradient of a scalar field can be defined in different ways.

13.2.2.1 Definition of the Gradient

The *gradient* of a function U is a vector $\operatorname{grad} U$, which can be assigned to every point P with the vector $\vec{\mathbf{r}}$ of a scalar field $U = U(\vec{\mathbf{r}})$, having the following properties:

1. The direction of $\operatorname{grad} U$ is always perpendicular to the direction of the level surface $U = \text{const}$, passing through the considered points P,

2. $\operatorname{grad} U$ points always into the direction in which the function U is increasing,

3. $|\operatorname{grad} U| = \dfrac{\partial U}{\partial \vec{\mathbf{n}}}$, i.e., the magnitude of $\operatorname{grad} U$ is equal to the directional derivative of U in the *normal direction*.

If the gradient is defined in another way, e.g., as a volume derivative or by the differential operator, then the previous defining properties became consequences of the definition.

13.2.2.2 Gradient and Directional Derivative

The directional derivative of the scalar field U with respect to the unit vector $\vec{\mathbf{c}}^{\,0}$ is equal to the projection of $\operatorname{grad} U$ onto the direction of the unit vector $\vec{\mathbf{c}}^{\,0}$:

$$\frac{\partial U}{\partial \vec{\mathbf{c}}^{\,0}} = \vec{\mathbf{c}}^{\,0} \cdot \operatorname{grad} U, \tag{13.34}$$

i.e., the directional derivative can be calculated as the dot product of the gradient and the unit vector pointing into the required direction.

Remark: The directional derivative at certain points in certain directions may also exist if the gradient does not exist there.

13.2.2.3 Gradient and Volume Derivative

The *gradient* U of the scalar field $U = U(\vec{\mathbf{r}})$ at a point $\vec{\mathbf{r}}$ can be defined as its *volume derivative*. If the following limit exists, then it is called the *gradient of* U at $\vec{\mathbf{r}}$:

$$\operatorname{grad} U = \lim_{V \to 0} \frac{\displaystyle\oiint_{(\Sigma)} U \, d\vec{\mathbf{S}}}{V}. \tag{13.35}$$

Here V is the volume of the region of space containing the point belonging to $\vec{\mathbf{r}}$ inside and bounded by the closed surface Σ. (If the independent variable is not a three-dimensional vector, then the gradient is defined by the differential operator.)

13.2.2.4 Further Properties of the Gradient

1. The absolute value of the gradient is greater if the level lines or level surfaces drawn as mentioned in 13.1.2.4, **2.**, p. 703, are more dense.

2. The gradient is the zero vector ($\operatorname{grad} U = \vec{\mathbf{0}}$) if U has a maximum or minimum at the considered point. The level lines or surfaces degenerate to a point there.

13.2.2.5 Gradient of the Scalar Field in Different Coordinates

1. Gradient in Cartesian Coordinates

$$\operatorname{grad} U = \frac{\partial U(x,y,z)}{\partial x}\vec{\mathbf{i}} + \frac{\partial U(x,y,z)}{\partial y}\vec{\mathbf{j}} + \frac{\partial U(x,y,z)}{\partial z}\vec{\mathbf{k}}. \tag{13.36}$$

2. Gradient in Cylindrical Coordinates ($x = \rho\cos\varphi,\ y = \rho\sin\varphi,\ z = z$)

$$\operatorname{grad} U = \operatorname{grad}_\rho U\vec{\mathbf{e}}_\rho + \operatorname{grad}_\varphi U\vec{\mathbf{e}}_\varphi + \operatorname{grad}_z U\vec{\mathbf{e}}_z \quad \text{with} \tag{13.37a}$$

$$\operatorname{grad}_\rho U = \frac{\partial U}{\partial \rho}, \quad \operatorname{grad}_\varphi U = \frac{1}{\rho}\frac{\partial U}{\partial \varphi}, \quad \operatorname{grad}_z U = \frac{\partial U}{\partial z}. \tag{13.37b}$$

3. Gradient in Spherical Coordinates ($x = r\sin\vartheta\cos\varphi,\ y = r\sin\vartheta\sin\varphi,\ z = r\cos\vartheta$)

$$\operatorname{grad} U = \operatorname{grad}_r U\vec{\mathbf{e}}_r + \operatorname{grad}_\vartheta U\vec{\mathbf{e}}_\vartheta + \operatorname{grad}_\varphi U\vec{\mathbf{e}}_\varphi \quad \text{with} \tag{13.38a}$$

$$\operatorname{grad}_r U = \frac{\partial U}{\partial r}, \quad \operatorname{grad}_\vartheta U = \frac{1}{r}\frac{\partial U}{\partial \vartheta}, \quad \operatorname{grad}_\varphi U = \frac{1}{r\sin\vartheta}\frac{\partial U}{\partial \varphi}. \tag{13.38b}$$

4. Gradient in General Orthogonal Coordinates (ξ, η, ζ)

For $\vec{\mathbf{r}}(\xi,\eta,\zeta) = x(\xi,\eta,\zeta)\vec{\mathbf{i}} + y(\xi,\eta,\zeta)\vec{\mathbf{j}} + z(\xi,\eta,\zeta)\vec{\mathbf{k}}$:

$$\operatorname{grad} U = \operatorname{grad}_\xi U\vec{\mathbf{e}}_\xi + \operatorname{grad}_\eta U\vec{\mathbf{e}}_\eta + \operatorname{grad}_\zeta U\vec{\mathbf{e}}_\zeta, \quad \text{where} \tag{13.39a}$$

$$\operatorname{grad}_\xi U = \frac{1}{\left|\frac{\partial \vec{\mathbf{r}}}{\partial \xi}\right|}\frac{\partial U}{\partial \xi}, \quad \operatorname{grad}_\eta U = \frac{1}{\left|\frac{\partial \vec{\mathbf{r}}}{\partial \eta}\right|}\frac{\partial U}{\partial \eta}, \quad \operatorname{grad}_\zeta U = \frac{1}{\left|\frac{\partial \vec{\mathbf{r}}}{\partial \zeta}\right|}\frac{\partial U}{\partial \zeta}. \tag{13.39b}$$

13.2.2.6 Rules of Calculations

Assuming in the followings that $\vec{\mathbf{c}}$ and c are constant, the following equalities hold:

$$\operatorname{grad} c = 0, \quad \operatorname{grad}(U_1 + U_2) = \operatorname{grad} U_1 + \operatorname{grad} U_2, \quad \operatorname{grad}(cU) = c\operatorname{grad} U. \tag{13.40}$$

$$\operatorname{grad}(U_1U_2) = U_1\operatorname{grad} U_2 + U_2\operatorname{grad} U_1, \quad \operatorname{grad}\varphi(U) = \frac{d\varphi}{dU}\operatorname{grad} U. \tag{13.41}$$

$$\operatorname{grad}(\vec{\mathbf{V}}_1 \cdot \vec{\mathbf{V}}_2) = (\vec{\mathbf{V}}_1 \cdot \operatorname{grad})\vec{\mathbf{V}}_2 + (\vec{\mathbf{V}}_2 \cdot \operatorname{grad})\vec{\mathbf{V}}_1 + \vec{\mathbf{V}}_1 \times \operatorname{rot}\vec{\mathbf{V}}_2 + \vec{\mathbf{V}}_2 \times \operatorname{rot}\vec{\mathbf{V}}_1. \tag{13.42}$$

$$\operatorname{grad}(\vec{\mathbf{r}} \cdot \vec{\mathbf{c}}) = \vec{\mathbf{c}}. \tag{13.43}$$

1. Differential of a Scalar Field as the Total Differential of the Function U

$$dU = \operatorname{grad} U \cdot d\vec{\mathbf{r}} = \frac{\partial U}{\partial x}dx + \frac{\partial U}{\partial y}dy + \frac{\partial U}{\partial z}dz. \tag{13.44}$$

2. Derivative of a Function U along a Space Curve $\vec{\mathbf{r}}(t)$

$$\frac{dU}{dt} = \frac{\partial U}{\partial x}\frac{dx}{dt} + \frac{\partial U}{\partial y}\frac{dy}{dt} + \frac{\partial U}{\partial z}\frac{dz}{dt}. \tag{13.45}$$

3. Gradient of a Central Field

$\operatorname{grad} U(r) = U'(r)\dfrac{\vec{\mathbf{r}}}{r}$ (spherical field), (13.46a) $\operatorname{grad} r = \dfrac{\vec{\mathbf{r}}}{r}$ (field of unit vectors). (13.46b)

13.2.3 Vector Gradient

The relation (13.32a) inspires the notation

$$\frac{\partial \vec{\mathbf{V}}}{\partial \vec{\mathbf{a}}} = \vec{\mathbf{a}} \cdot \operatorname{grad}(V_x\vec{\mathbf{e}}_x + V_y\vec{\mathbf{e}}_y + V_z\vec{\mathbf{e}}_z) = \vec{\mathbf{a}} \cdot \operatorname{grad}\vec{\mathbf{V}} \tag{13.47a}$$

where grad $\vec{\mathbf{V}}$ is called the *vector gradient*. It follows from the matrix notation of (13.47a) that the vector gradient, as a tensor, can be represented by a matrix:

$$(\vec{\mathbf{a}}\cdot\operatorname{grad})\vec{\mathbf{V}} = \begin{pmatrix} \frac{\partial V_x}{\partial x} & \frac{\partial V_x}{\partial y} & \frac{\partial V_x}{\partial z} \\ \frac{\partial V_y}{\partial x} & \frac{\partial V_y}{\partial y} & \frac{\partial V_y}{\partial z} \\ \frac{\partial V_z}{\partial x} & \frac{\partial V_z}{\partial y} & \frac{\partial V_z}{\partial z} \end{pmatrix} \begin{pmatrix} a_x \\ a_y \\ a_z \end{pmatrix}, \quad (13.47b) \qquad \operatorname{grad}\vec{\mathbf{V}} = \begin{pmatrix} \frac{\partial V_x}{\partial x} & \frac{\partial V_x}{\partial y} & \frac{\partial V_x}{\partial z} \\ \frac{\partial V_y}{\partial x} & \frac{\partial V_y}{\partial y} & \frac{\partial V_y}{\partial z} \\ \frac{\partial V_z}{\partial x} & \frac{\partial V_z}{\partial y} & \frac{\partial V_z}{\partial z} \end{pmatrix}. \quad (13.47c)$$

These types of tensors have a very important role in engineering sciences, e.g., for the description of tension and elasticity (see 4.3.2, **4.**, p. 282.

13.2.4 Divergence of Vector Fields

13.2.4.1 Definition of Divergence

To a vector field $\vec{\mathbf{V}}(\vec{\mathbf{r}})$ a scalar field can be assigned which is called its *divergence*. The divergence is defined as a space derivative of the vector field at a point $\vec{\mathbf{r}}$:

$$\operatorname{div}\vec{\mathbf{V}} = \lim_{V\to 0} \frac{\oiint\limits_{(\Sigma)} \vec{\mathbf{V}}\cdot d\vec{\mathbf{S}}}{V}. \quad (13.48)$$

If the vector field $\vec{\mathbf{V}}$ is considered as a stream field, then the divergence can be considered as the fluid output or source, because it gives the amount of fluid given in a unit of volume during a unit of time flowing by the considered point of the vector field $\vec{\mathbf{V}}$. In the case div $\vec{\mathbf{V}} > 0$ the point is called a *source*, in the case div $\vec{\mathbf{V}} < 0$ it is called a *sink*.

13.2.4.2 Divergence in Different Coordinates

1. Divergence in Cartesian Coordinates

$$\operatorname{div}\vec{\mathbf{V}} = \frac{\partial V_x}{\partial x} + \frac{\partial V_y}{\partial y} + \frac{\partial V_z}{\partial z} \quad (13.49a) \qquad \text{with} \quad \vec{\mathbf{V}}(x,y,z) = V_x\vec{\mathbf{i}} + V_y\vec{\mathbf{j}} + V_z\vec{\mathbf{k}}. \quad (13.49b)$$

The scalar field div $\vec{\mathbf{V}}$ can be represented as the dot product of the nabla operator ∇ and the vector $\vec{\mathbf{V}}$ as

$$\operatorname{div}\vec{\mathbf{V}} = \nabla\cdot\vec{\mathbf{V}} \quad (13.49c)$$

and it is translation and rotation invariant, i.e., scalar invariant (see 4.3.3.2, p. 283).

2. Divergence in Cylindrical Coordinates

$$\operatorname{div}\vec{\mathbf{V}} = \frac{1}{\rho}\frac{\partial(\rho V_\rho)}{\partial\rho} + \frac{1}{\rho}\frac{\partial V_\varphi}{\partial\varphi} + \frac{\partial V_z}{\partial z} \quad (13.50a) \quad \text{with} \quad \vec{\mathbf{V}}(\rho,\varphi,z) = V_\rho\vec{\mathbf{e}}_\rho + V_\varphi\vec{\mathbf{e}}_\varphi + V_z\vec{\mathbf{e}}_z. \quad (13.50b)$$

3. Divergence in Spherical Coordinates

$$\operatorname{div}\vec{\mathbf{V}} = \frac{1}{r^2}\frac{\partial(r^2 V_r)}{\partial r} + \frac{1}{r\sin\vartheta}\frac{\partial(\sin\vartheta V_\vartheta)}{\partial\vartheta} + \frac{1}{r\sin\vartheta}\frac{\partial V_\varphi}{\partial\varphi} \quad (13.51a)$$

$$\text{with} \quad \vec{\mathbf{V}}(r,\vartheta,\varphi) = V_r\vec{\mathbf{e}}_r + V_\vartheta\vec{\mathbf{e}}_\vartheta + V_\varphi\vec{\mathbf{e}}_\varphi. \quad (13.51b)$$

4. Divergence in General Orthogonal Coordinates

$$\operatorname{div}\vec{\mathbf{V}} = \frac{1}{\mathrm{D}}\left\{\frac{\partial}{\partial\xi}\left(\left|\frac{\partial\vec{\mathbf{r}}}{\partial\eta}\right|\left|\frac{\partial\vec{\mathbf{r}}}{\partial\zeta}\right|V_\xi\right) + \frac{\partial}{\partial\eta}\left(\left|\frac{\partial\vec{\mathbf{r}}}{\partial\zeta}\right|\left|\frac{\partial\vec{\mathbf{r}}}{\partial\xi}\right|V_\eta\right) + \frac{\partial}{\partial\zeta}\left(\left|\frac{\partial\vec{\mathbf{r}}}{\partial\xi}\right|\left|\frac{\partial\vec{\mathbf{r}}}{\partial\eta}\right|V_\zeta\right)\right\} \quad (13.52a)$$

with $\vec{\mathbf{r}}(\xi,\eta,\zeta) = x(\xi,\eta,\zeta)\vec{\mathbf{i}} + y(\xi,\eta,\zeta)\vec{\mathbf{j}} + z(\xi,\eta,\zeta)\vec{\mathbf{k}},$ (13.52b)

$$\mathrm{D} = \left|\left(\frac{\partial\vec{\mathbf{r}}}{\partial\xi}\frac{\partial\vec{\mathbf{r}}}{\partial\eta}\frac{\partial\vec{\mathbf{r}}}{\partial\zeta}\right)\right| = \left|\frac{\partial\vec{\mathbf{r}}}{\partial\xi}\right|\cdot\left|\frac{\partial\vec{\mathbf{r}}}{\partial\eta}\right|\cdot\left|\frac{\partial\vec{\mathbf{r}}}{\partial\zeta}\right| \quad (13.52c) \text{ and } \vec{\mathbf{V}}(\xi,\eta,\zeta) = V_\xi\vec{\mathbf{e}}_\xi + V_\eta\vec{\mathbf{e}}_\eta + V_\zeta\vec{\mathbf{e}}_\zeta. \quad (13.52d)$$

13.2.4.3 Rules for Evaluation of the Divergence

$$\operatorname{div}\vec{\mathbf{c}} = 0, \quad \operatorname{div}(\vec{\mathbf{V}}_1 + \vec{\mathbf{V}}_2) = \operatorname{div}\vec{\mathbf{V}}_1 + \operatorname{div}\vec{\mathbf{V}}_2, \quad \operatorname{div}(c\vec{\mathbf{V}}) = c\operatorname{div}\vec{\mathbf{V}}. \quad (13.53)$$

$$\operatorname{div}(U\vec{\mathbf{V}}) = U\operatorname{div}\vec{\mathbf{V}} + \vec{\mathbf{V}}\cdot\operatorname{grad}U \qquad \left(\text{especially } \operatorname{div}(r\vec{\mathbf{c}}) = \frac{\vec{\mathbf{r}}\cdot\vec{\mathbf{c}}}{r}\right). \quad (13.54)$$

$$\operatorname{div}(\vec{\mathbf{V}}_1\times\vec{\mathbf{V}}_2) = \vec{\mathbf{V}}_2\cdot\operatorname{rot}\vec{\mathbf{V}}_1 - \vec{\mathbf{V}}_1\cdot\operatorname{rot}\vec{\mathbf{V}}_2\,. \quad (13.55)$$

13.2.4.4 Divergence of a Central Field

$$\operatorname{div}\vec{\mathbf{r}} = 3, \quad \operatorname{div}\varphi(r)\vec{\mathbf{r}} = 3\varphi(r) + r\varphi'(r). \quad (13.56)$$

13.2.5 Rotation of Vector Fields

13.2.5.1 Definitions of the Rotation

1. Definition

The *rotation* or *curl* of a vector field $\vec{\mathbf{V}}$ at the point $\vec{\mathbf{r}}$ is a vector denoted by $\operatorname{rot}\vec{\mathbf{V}}$, $\operatorname{curl}\vec{\mathbf{V}}$ or with the nabla operator $\nabla\times\vec{\mathbf{V}}$, and defined as the negative space derivative of the vector field:

$$\operatorname{rot}\vec{\mathbf{V}} = -\lim_{V\to 0}\frac{\displaystyle\oiint_{(\Sigma)}\vec{\mathbf{V}}\times d\vec{\mathbf{S}}}{V} = \lim_{V\to 0}\frac{\displaystyle\oiint_{(\Sigma)} d\vec{\mathbf{S}}\times\vec{\mathbf{V}}}{V}. \quad (13.57)$$

2. Definition

The vector field of the rotation of the vector field $\vec{\mathbf{V}}(\vec{\mathbf{r}})$ can be defined in the following way:

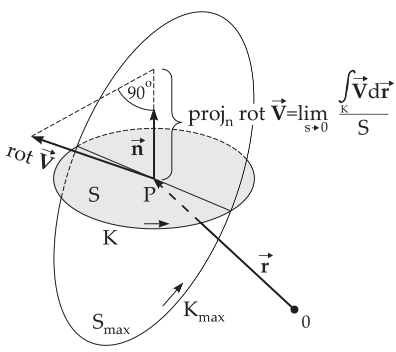


Figure 13.12

a) Putting a small surface sheet S **(Fig. 13.12)** through the point $\vec{\mathbf{r}}$ and describing this surface sheet by a vector $\vec{\mathbf{S}}$ whose direction is the direction of the surface normal $\vec{\mathbf{n}}$ and its absolute value is equal to the area of this surface region. The boundary of this surface is denoted by C.

b) Evaluating the integral $\displaystyle\oint_{(C)}\vec{\mathbf{V}}\cdot d\vec{\mathbf{r}}$ along the closed boundary curve C of the surface (the sense of the curve is positive looking to the surface from the direction of the surface normal (see **Fig. 13.12**).

c) Determining the limit (if it exists) $\displaystyle\lim_{S\to 0}\frac{1}{S}\oint_{(C)}\vec{\mathbf{V}}\cdot d\vec{\mathbf{r}}$, while the position of the surface sheet remains unchanged.

d) Changing the position of the surface sheet in order to get a maximum value of the limit. The surface area in this position is $S_{\max}$ and the corresponding boundary curve is $C_{\max}$.

e) Determining the vector $\operatorname{rot}\vec{\mathbf{r}}$ at the point $\vec{\mathbf{r}}$, whose absolute value is equal to the maximum value

found above and its direction coincides with the direction of the surface normal of the corresponding surface. Then one gets:

$$\left|\operatorname{rot}\vec{\mathbf{V}}\right| = \lim_{S_{\max}\to 0} \frac{\oint\limits_{(C_{\max})} \vec{\mathbf{V}}\cdot d\vec{\mathbf{r}}}{S_{\max}}. \tag{13.58a}$$

The projection of rot $\vec{\mathbf{V}}$ onto the surface normal $\vec{\mathbf{n}}$ of a surface with area S, i.e., the component of the vector rot $\vec{\mathbf{V}}$ in an arbitrary direction $\vec{\mathbf{n}} = \vec{\mathbf{l}}$ is

$$\vec{\mathbf{l}}\cdot\operatorname{rot}\vec{\mathbf{V}} = \operatorname{rot}_l\vec{\mathbf{V}} = \lim_{S\to 0}\frac{\oint\limits_{(C)} \vec{\mathbf{V}}\cdot d\vec{\mathbf{r}}}{S}. \tag{13.58b}$$

The vector lines of the field rot $\vec{\mathbf{V}}$ are called the *curl lines of the vector field* $\vec{\mathbf{V}}$.

13.2.5.2 Rotation in Different Coordinates

1. Rotation in Cartesian Coordinates

$$\operatorname{rot}\vec{\mathbf{V}} = \vec{\mathbf{i}}\left(\frac{\partial V_z}{\partial y} - \frac{\partial V_y}{\partial z}\right) + \vec{\mathbf{j}}\left(\frac{\partial V_x}{\partial z} - \frac{\partial V_z}{\partial x}\right) + \vec{\mathbf{k}}\left(\frac{\partial V_y}{\partial x} - \frac{\partial V_x}{\partial y}\right) = \begin{vmatrix} \vec{\mathbf{i}} & \vec{\mathbf{j}} & \vec{\mathbf{k}} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ V_x & V_y & V_z \end{vmatrix}. \tag{13.59a}$$

The vector field rot $\vec{\mathbf{V}}$ can be represented as the cross product of the nabla operator ∇ and the vector $\vec{\mathbf{V}}$:

$$\operatorname{rot}\vec{\mathbf{V}} = \nabla\times\vec{\mathbf{V}}. \tag{13.59b}$$

2. Rotation in Cylindrical Coordinates

$$\operatorname{rot}\vec{\mathbf{V}} = \operatorname{rot}_\rho\vec{\mathbf{V}}\vec{\mathbf{e}}_\rho + \operatorname{rot}_\varphi\vec{\mathbf{V}}\vec{\mathbf{e}}_\varphi + \operatorname{rot}_z\vec{\mathbf{V}}\vec{\mathbf{e}}_z \quad \text{with} \tag{13.60a}$$

$$\operatorname{rot}_\rho\vec{\mathbf{V}} = \frac{1}{\rho}\frac{\partial V_z}{\partial\varphi} - \frac{\partial V_\varphi}{\partial z}, \quad \operatorname{rot}_\varphi\vec{\mathbf{V}} = \frac{\partial V_\rho}{\partial z} - \frac{\partial V_z}{\partial\rho}, \quad \operatorname{rot}_z\vec{\mathbf{V}} = \frac{1}{\rho}\left\{\frac{\partial}{\partial\rho}(\rho V_\varphi) - \frac{\partial V_\rho}{\partial\varphi}\right\}. \tag{13.60b}$$

3. Rotation in Spherical Coordinates

$$\operatorname{rot}\vec{\mathbf{V}} = \operatorname{rot}_r\vec{\mathbf{V}}\vec{\mathbf{e}}_r + \operatorname{rot}_\vartheta\vec{\mathbf{V}}\vec{\mathbf{e}}_\vartheta + \operatorname{rot}_\varphi\vec{\mathbf{V}}\vec{\mathbf{e}}_\varphi \quad \text{with} \tag{13.61a}$$

$$\left.\begin{aligned} \operatorname{rot}_r\vec{\mathbf{V}} &= \frac{1}{r\sin\vartheta}\left\{\frac{\partial}{\partial\vartheta}(\sin\vartheta V_\varphi) - \frac{\partial V_\vartheta}{\partial\varphi}\right\}, \\ \operatorname{rot}_\vartheta\vec{\mathbf{V}} &= \frac{1}{r\sin\vartheta}\frac{\partial V_r}{\partial\varphi} - \frac{1}{r}\frac{\partial}{\partial r}(rV_\varphi), \\ \operatorname{rot}_\varphi\vec{\mathbf{V}} &= \frac{1}{r}\left\{\frac{\partial}{\partial r}(rV_\vartheta) - \frac{\partial V_r}{\partial\vartheta}\right\}. \end{aligned}\right\} \tag{13.61b}$$

4. Rotation in General Orthogonal Coordinates

$$\operatorname{rot}\vec{\mathbf{V}} = \operatorname{rot}_\xi\vec{\mathbf{V}}\vec{\mathbf{e}}_\xi + \operatorname{rot}_\eta\vec{\mathbf{V}}\vec{\mathbf{e}}_\eta + \operatorname{rot}_\zeta\vec{\mathbf{V}}\vec{\mathbf{e}}_\zeta \quad \text{with} \tag{13.62a}$$

$$\left.\begin{aligned}
\operatorname{rot}_\xi \vec{\mathbf{V}} &= \frac{1}{\mathrm{D}}\left|\frac{\partial\vec{\mathbf{r}}}{\partial\xi}\right|\left[\frac{\partial}{\partial\eta}\left(\left|\frac{\partial\vec{\mathbf{r}}}{\partial\zeta}\right|V_\zeta\right)-\frac{\partial}{\partial\zeta}\left(\left|\frac{\partial\vec{\mathbf{r}}}{\partial\eta}\right|V_\eta\right)\right],\\
\operatorname{rot}_\eta \vec{\mathbf{V}} &= \frac{1}{\mathrm{D}}\left|\frac{\partial\vec{\mathbf{r}}}{\partial\eta}\right|\left[\frac{\partial}{\partial\zeta}\left(\left|\frac{\partial\vec{\mathbf{r}}}{\partial\xi}\right|V_\xi\right)-\frac{\partial}{\partial\xi}\left(\left|\frac{\partial\vec{\mathbf{r}}}{\partial\zeta}\right|V_\zeta\right)\right],\\
\operatorname{rot}_\zeta \vec{\mathbf{V}} &= \frac{1}{\mathrm{D}}\left|\frac{\partial\vec{\mathbf{r}}}{\partial\zeta}\right|\left[\frac{\partial}{\partial\xi}\left(\left|\frac{\partial\vec{\mathbf{r}}}{\partial\eta}\right|V_\eta\right)-\frac{\partial}{\partial\eta}\left(\left|\frac{\partial\vec{\mathbf{r}}}{\partial\xi}\right|V_\xi\right)\right],
\end{aligned}\right\} \tag{13.62b}$$

$$\vec{\mathbf{r}}(\xi,\eta,\zeta) = x(\xi,\eta,\zeta)\vec{\mathbf{i}} + y(\xi,\eta,\zeta)\vec{\mathbf{j}} + z(\xi,\eta,\zeta)\vec{\mathbf{k}}; \qquad \mathrm{D} = \left|\frac{\partial\vec{\mathbf{r}}}{\partial\xi}\right|\cdot\left|\frac{\partial\vec{\mathbf{r}}}{\partial\eta}\right|\cdot\left|\frac{\partial\vec{\mathbf{r}}}{\partial\zeta}\right|. \tag{13.62c}$$

13.2.5.3 Rules for Evaluating the Rotation

$$\operatorname{rot}(\vec{\mathbf{V}}_1+\vec{\mathbf{V}}_2) = \operatorname{rot}\vec{\mathbf{V}}_1 + \operatorname{rot}\vec{\mathbf{V}}_2, \quad \operatorname{rot}(c\vec{\mathbf{V}}) = c\operatorname{rot}\vec{\mathbf{V}}. \tag{13.63}$$

$$\operatorname{rot}(U\vec{\mathbf{V}}) = U\operatorname{rot}\vec{\mathbf{V}} + \operatorname{grad}U\times\vec{\mathbf{V}}. \tag{13.64}$$

$$\operatorname{rot}(\vec{\mathbf{V}}_1\times\vec{\mathbf{V}}_2) = (\vec{\mathbf{V}}_2\cdot\operatorname{grad})\vec{\mathbf{V}}_1 - (\vec{\mathbf{V}}_1\cdot\operatorname{grad})\vec{\mathbf{V}}_2 + \vec{\mathbf{V}}_1\operatorname{div}\vec{\mathbf{V}}_2 - \vec{\mathbf{V}}_2\operatorname{div}\vec{\mathbf{V}}_1. \tag{13.65}$$

13.2.5.4 Rotation of a Potential Field

This also follows from the Stokes theorem (see 13.3.3.2, p. 725) that the rotation of a potential field is identically zero:

$$\operatorname{rot}\vec{\mathbf{V}} = \operatorname{rot}(\operatorname{grad}\mathrm{U}) = \vec{\mathbf{0}}. \tag{13.66}$$

This also follows from (13.59a) for $\vec{\mathbf{V}} = \operatorname{grad}U$, if the assumptions of the Schwarz interchanging theorem are fulfilled (see 6.2.2.2, **1.**, p. 448).

■ For $\vec{\mathbf{r}} = x\vec{\mathbf{i}} + y\vec{\mathbf{j}} + z\vec{\mathbf{k}}$ with $r = |\vec{\mathbf{r}}| = \sqrt{x^2+y^2+z^2}$ holds: $\operatorname{rot}\vec{\mathbf{r}} = \vec{\mathbf{0}}$ and $\operatorname{rot}(\varphi(r)\vec{\mathbf{r}}) = \vec{\mathbf{0}}$, where $\varphi(r)$ is a differentiable function of r.

13.2.6 Nabla Operator, Laplace Operator

13.2.6.1 Nabla Operator

The symbolic vector ∇ is called the *nabla operator*. Its use simplifies the representation of and calculations with space differential operators. In Cartesian coordinates holds

$$\nabla = \frac{\partial}{\partial x}\vec{\mathbf{i}} + \frac{\partial}{\partial y}\vec{\mathbf{j}} + \frac{\partial}{\partial z}\vec{\mathbf{k}}. \tag{13.67}$$

The components of the nabla operator are considered as partial differential operators, i.e., the symbol $\frac{\partial}{\partial x}$ means partial differentiation with respect to x, where the other variables are considered as constants.

The formulas for *spatial differential operators* in Cartesian coordinates can be obtained by formal multiplication of this vector operator by the scalar U or by the vector $\vec{\mathbf{V}}$. For instance, in the case of the operators *gradient, vector gradient, divergence*, and *rotation*:

$$\operatorname{grad}U = \nabla U \qquad \text{(gradient of } U \quad \text{(see 13.2.2, p. 710))}, \tag{13.68a}$$

$$\operatorname{grad}\vec{\mathbf{V}} = \nabla\vec{\mathbf{V}} \qquad \text{(vector gradient of } \vec{\mathbf{V}} \quad \text{(see 13.2.3, p. 711))}, \tag{13.68b}$$

$$\operatorname{div}\vec{\mathbf{V}} = \nabla\cdot\vec{\mathbf{V}} \qquad \text{(divergence of } \vec{\mathbf{V}} \quad \text{(see 13.2.4, p. 712))}, \tag{13.68c}$$

$$\operatorname{rot}\vec{\mathbf{V}} = \nabla\times\vec{\mathbf{V}} \qquad \text{(rotation or curl of } \vec{\mathbf{V}} \quad \text{(see 13.2.5, p. 713))}. \tag{13.68d}$$

13.2.6.2 Rules for Calculations with the Nabla Operator

1. If ∇ stands in front of a linear combination $\sum a_i X_i$ with constants a_i and with point functions X_i, then, independently of whether they are scalar or vector functions, we have the formula:

$$\nabla(\sum a_i X_i) = \sum a_i \nabla X_i. \tag{13.69}$$

2. If ∇ is applied to a product of scalar or vector functions, then it has to be applied to each of these functions after each other and the results are to be added. There is a $\downarrow$ above the symbol of the function submitted to the operation

$$\nabla(XYZ) = \nabla(\overset{\downarrow}{X}YZ) + \nabla(X\overset{\downarrow}{Y}Z) + \nabla(XY\overset{\downarrow}{Z}), \quad \text{i.e.,} \tag{13.70}$$

$$\nabla(XYZ) = (\nabla X)YZ + X(\nabla Y)Z) + XY(\nabla Z).$$

Then the products have to be transformed according to vector algebra so as the operator ∇ is applied to only one factor with the sign $\downarrow$. Having performed the computation one omits that sign.

■ **A:** $\operatorname{div}(U\vec{\mathbf{V}}) = \nabla(U\vec{\mathbf{V}}) = \nabla(\overset{\downarrow}{U}\vec{\mathbf{V}}) + \nabla(U\overset{\downarrow}{\vec{\mathbf{V}}}) = \vec{\mathbf{V}} \cdot \nabla U + U\nabla \cdot \vec{\mathbf{V}} = \vec{\mathbf{V}} \cdot \operatorname{grad} U + U \operatorname{div} \vec{\mathbf{V}}.$

■ **B:** $\operatorname{grad}(\vec{\mathbf{V}}_1\vec{\mathbf{V}}_2) = \nabla(\vec{\mathbf{V}}_1\vec{\mathbf{V}}_2) = \nabla(\overset{\downarrow}{\vec{\mathbf{V}}_1}\vec{\mathbf{V}}_2) + \nabla(\vec{\mathbf{V}}_1\overset{\downarrow}{\vec{\mathbf{V}}_2})$. Because $\vec{\mathbf{b}}(\vec{\mathbf{a}}\vec{\mathbf{c}}) = (\vec{\mathbf{a}}\vec{\mathbf{b}})\vec{\mathbf{c}} + \vec{\mathbf{a}} \times (\vec{\mathbf{b}} \times \vec{\mathbf{c}})$ follows: $\operatorname{grad}(\vec{\mathbf{V}}_1\vec{\mathbf{V}}_2) = (\vec{\mathbf{V}}_2\nabla)\vec{\mathbf{V}}_1 + \vec{\mathbf{V}}_2 \times (\nabla \times \vec{\mathbf{V}}_1) + (\vec{\mathbf{V}}_1\nabla)\vec{\mathbf{V}}_2 + \vec{\mathbf{V}}_1 \times (\nabla \times \vec{\mathbf{V}}_2)$

$= (\vec{\mathbf{V}}_2 \operatorname{grad})\vec{\mathbf{V}}_1 + \vec{\mathbf{V}}_2 \times \operatorname{rot} \vec{\mathbf{V}}_1 + (\vec{\mathbf{V}}_1 \operatorname{grad})\vec{\mathbf{V}}_2 + \vec{\mathbf{V}}_1 \times \operatorname{rot} \vec{\mathbf{V}}_2.$

13.2.6.3 Vector Gradient

The vector gradient $\operatorname{grad} \vec{\mathbf{V}}$ is represented by the nabla operator as

$$\operatorname{grad} \vec{\mathbf{V}} = \nabla\vec{\mathbf{V}}. \tag{13.71a}$$

The expression occurring in the vector gradient $(\vec{\mathbf{a}} \cdot \nabla)\vec{\mathbf{V}}$ (see (13.32b), p. 709) has the form:

$$2(\vec{\mathbf{a}} \cdot \nabla)\vec{\mathbf{V}} = \operatorname{rot}(\vec{\mathbf{V}} \times \vec{\mathbf{a}}) + \operatorname{grad}(\vec{\mathbf{a}}\vec{\mathbf{V}}) + \vec{\mathbf{a}} \operatorname{div} \vec{\mathbf{V}} - \vec{\mathbf{V}} \operatorname{div} \vec{\mathbf{a}} - \vec{\mathbf{a}} \times \operatorname{rot} \vec{\mathbf{V}} - \vec{\mathbf{V}} \times \operatorname{rot} \vec{\mathbf{a}}. \tag{13.71b}$$

In particular one gets for $\vec{\mathbf{r}} = x\vec{\mathbf{i}} + y\vec{\mathbf{j}} + z\vec{\mathbf{k}}$:

$$(\vec{\mathbf{a}} \cdot \nabla)\vec{\mathbf{r}} = \vec{\mathbf{a}}. \tag{13.71c}$$

13.2.6.4 Nabla Operator Applied Twice

For every field $\vec{\mathbf{V}}$:

$$\nabla(\nabla \times \vec{\mathbf{V}}) = \operatorname{div} \operatorname{rot} \vec{\mathbf{V}} \equiv 0, \tag{13.72}$$

$$\nabla \times (\nabla U) = \operatorname{rot} \operatorname{grad} U \equiv \vec{\mathbf{0}}, \tag{13.73}$$

$$\nabla(\nabla U) = \operatorname{div} \operatorname{grad} U = \Delta U. \tag{13.74}$$

13.2.6.5 Laplace Operator

1. Definition

The dot product of the nabla operator with itself is called the *Laplace operator*:

$$\Delta = \nabla \cdot \nabla = \nabla^2. \tag{13.75}$$

The Laplace operator is not a vector. It prescribes the summation of the second partial derivatives. It can be applied to scalar functions as well as to vector functions. The application to a vector function, componentwise, results in a vector.

The Laplace operator is an *invariant*, i.e., it does not change during translation and/or rotation of the coordinate system.

2. Formulas for the Laplace Operator in Different Coordinates

Here the Laplace operator is applied to the scalar point function $U(\vec{\mathbf{r}})$. Then the result is a scalar. The application of it for vector functions $\vec{\mathbf{V}}(\vec{\mathbf{r}})$ results in a vector $\Delta\vec{\mathbf{V}}$ with components ΔV_x, ΔV_y, ΔV_z.

1. Laplace Operator in Cartesian Coordinates

$$\Delta U(x,y,z) = \frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} + \frac{\partial^2 U}{\partial z^2}. \tag{13.76}$$

2. Laplace Operator in Cylindrical Coordinates

$$\Delta U(\rho,\varphi,z) = \frac{1}{\rho}\frac{\partial}{\partial\rho}\left(\rho\frac{\partial U}{\partial\rho}\right) + \frac{1}{\rho^2}\frac{\partial^2 U}{\partial\varphi^2} + \frac{\partial^2 U}{\partial z^2}. \tag{13.77}$$

3. Laplace Operator in Spherical Coordinates

$$\Delta U(r,\vartheta,\varphi) = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial U}{\partial r}\right) + \frac{1}{r^2\sin\vartheta}\frac{\partial}{\partial\vartheta}\left(\sin\vartheta\frac{\partial U}{\partial\vartheta}\right) + \frac{1}{r^2\sin^2\vartheta}\frac{\partial^2 U}{\partial\varphi^2}. \tag{13.78}$$

4. Laplace Operator in General Orthogonal Coordinates

$$\Delta U(\xi,\eta,\zeta) = \frac{1}{\mathrm{D}}\left[\frac{\partial}{\partial\xi}\left(\frac{\mathrm{D}}{\left|\frac{\partial\vec{\mathbf{r}}}{\partial\xi}\right|^2}\frac{\partial U}{\partial\xi}\right) + \frac{\partial}{\partial\eta}\left(\frac{\mathrm{D}}{\left|\frac{\partial\vec{\mathbf{r}}}{\partial\eta}\right|^2}\frac{\partial U}{\partial\eta}\right) + \frac{\partial}{\partial\zeta}\left(\frac{\mathrm{D}}{\left|\frac{\partial\vec{\mathbf{r}}}{\partial\zeta}\right|^2}\frac{\partial U}{\partial\zeta}\right)\right] \quad \text{with} \tag{13.79a}$$

$$\vec{\mathbf{r}}(\xi,\eta,\zeta) = x(\xi,\eta,\zeta)\vec{\mathbf{i}} + y(\xi,\eta,\zeta)\vec{\mathbf{j}} + z(\xi,\eta,\zeta)\vec{\mathbf{k}}, \tag{13.79b}$$

$$\mathrm{D} = \left|\frac{\partial\vec{\mathbf{r}}}{\partial\xi}\right|\cdot\left|\frac{\partial\vec{\mathbf{r}}}{\partial\eta}\right|\cdot\left|\frac{\partial\vec{\mathbf{r}}}{\partial\zeta}\right|. \tag{13.79c}$$

3. Special Relations between the Nabla Operator and Laplace Operator

$$\nabla(\nabla\cdot\vec{\mathbf{V}}) = \operatorname{grad}\operatorname{div}\vec{\mathbf{V}}, \tag{13.80}$$

$$\nabla\times(\nabla\times\vec{\mathbf{V}}) = \operatorname{rot}\operatorname{rot}\vec{\mathbf{V}}, \tag{13.81}$$

$$\nabla(\nabla\cdot\vec{\mathbf{V}}) - \nabla\times(\nabla\times\vec{\mathbf{V}}) = \Delta\vec{\mathbf{V}}, \quad \text{where} \tag{13.82}$$

$$\Delta\vec{\mathbf{V}} = (\nabla\cdot\nabla)\vec{\mathbf{V}} = \Delta V_x\vec{\mathbf{i}} + \Delta V_y\vec{\mathbf{j}} + \Delta V_z\vec{\mathbf{k}} = \left(\frac{\partial^2 V_x}{\partial x^2} + \frac{\partial^2 V_x}{\partial y^2} + \frac{\partial^2 V_x}{\partial z^2}\right)\vec{\mathbf{i}}$$
$$+\left(\frac{\partial^2 V_y}{\partial x^2} + \frac{\partial^2 V_y}{\partial y^2} + \frac{\partial^2 V_y}{\partial z^2}\right)\vec{\mathbf{j}} + \left(\frac{\partial^2 V_z}{\partial x^2} + \frac{\partial^2 V_z}{\partial y^2} + \frac{\partial^2 V_z}{\partial z^2}\right)\vec{\mathbf{k}}. \tag{13.83}$$

13.2.7 Review of Spatial Differential Operations

13.2.7.1 Rules of Calculation for Spatial Differential Operators

U, U_1, U_2 and F are scalar functions; c is a constant; $\vec{\mathbf{V}}, \vec{\mathbf{V}}_1, \vec{\mathbf{V}}_2$ are vector functions:

$$\operatorname{grad}(U_1+U_2) = \operatorname{grad}U_1 + \operatorname{grad}U_2. \tag{13.84}$$

$$\operatorname{grad}(cU) = c\operatorname{grad}U. \tag{13.85}$$

$$\operatorname{grad}(U_1U_2) = U_1\operatorname{grad}U_2 + U_2\operatorname{grad}U_1. \tag{13.86}$$

$$\operatorname{grad}F(U) = F'(U)\operatorname{grad}U. \tag{13.87}$$

$$\operatorname{div}(\vec{\mathbf{V}}_1+\vec{\mathbf{V}}_2) = \operatorname{div}\vec{\mathbf{V}}_1 + \operatorname{div}\vec{\mathbf{V}}_2. \tag{13.88}$$

$$\operatorname{div}(c\vec{\mathbf{V}}) = c\operatorname{div}\vec{\mathbf{V}}. \tag{13.89}$$

$$\operatorname{div}(U\vec{\mathbf{V}}) = \vec{\mathbf{V}} \cdot \operatorname{grad} U + U \operatorname{div} \vec{\mathbf{V}}. \quad (13.90)$$

$$\operatorname{rot}(\vec{\mathbf{V}}_1 + \vec{\mathbf{V}}_2) = \operatorname{rot} \vec{\mathbf{V}}_1 + \operatorname{rot} \vec{\mathbf{V}}_2. \quad (13.91)$$

$$\operatorname{rot}(c\vec{\mathbf{V}}) = c \operatorname{rot} \vec{\mathbf{V}}. \quad (13.92)$$

$$\operatorname{rot}(U\vec{\mathbf{V}}) = U \operatorname{rot} \vec{\mathbf{V}} - \vec{\mathbf{V}} \times \operatorname{grad} U. \quad (13.93)$$

$$\operatorname{div} \operatorname{rot} \vec{\mathbf{V}} \equiv 0. \quad (13.94)$$

$$\operatorname{rot} \operatorname{grad} U \equiv \vec{\mathbf{0}} \qquad \text{(zero vector)}. \quad (13.95)$$

$$\operatorname{div} \operatorname{grad} U = \Delta U. \quad (13.96)$$

$$\operatorname{rot} \operatorname{rot} \vec{\mathbf{V}} = \operatorname{grad} \operatorname{div} \vec{\mathbf{V}} - \Delta \vec{\mathbf{V}}. \quad (13.97)$$

$$\operatorname{div}(\vec{\mathbf{V}}_1 \times \vec{\mathbf{V}}_2) = \vec{\mathbf{V}}_2 \cdot \operatorname{rot} \vec{\mathbf{V}}_1 - \vec{\mathbf{V}}_1 \cdot \operatorname{rot} \vec{\mathbf{V}}_2. \quad (13.98)$$

13.2.7.2 Expressions of Vector Analysis in Cartesian, Cylindrical, and Spherical Coordinates (see Table 13.2)

Table 13.2 Expressions of vector analysis in Cartesian, cylindrical, and spherical coordinates

	Cartesian coordinates	Cylindrical coordinates	Spherical coordinates
$d\vec{\mathbf{s}} = d\vec{\mathbf{r}}$	$\vec{\mathbf{e}}_x dx + \vec{\mathbf{e}}_y dy + \vec{\mathbf{e}}_z dz$	$\vec{\mathbf{e}}_\rho d\rho + \vec{\mathbf{e}}_\varphi \rho d\varphi + \vec{\mathbf{e}}_z dz$	$\vec{\mathbf{e}}_r dr + \vec{\mathbf{e}}_\vartheta r d\vartheta + \vec{\mathbf{e}}_\varphi r \sin\vartheta d\varphi$
grad$U$	$\vec{\mathbf{e}}_x \frac{\partial U}{\partial x} + \vec{\mathbf{e}}_y \frac{\partial U}{\partial y} + \vec{\mathbf{e}}_z \frac{\partial U}{\partial z}$	$\vec{\mathbf{e}}_\rho \frac{\partial U}{\partial \rho} + \vec{\mathbf{e}}_\varphi \frac{1}{\rho}\frac{\partial U}{\partial \varphi} + \vec{\mathbf{e}}_z \frac{\partial U}{\partial z}$	$\vec{\mathbf{e}}_r \frac{\partial U}{\partial r} + \vec{\mathbf{e}}_\vartheta \frac{1}{r}\frac{\partial U}{\partial \vartheta} + \vec{\mathbf{e}}_\varphi \frac{1}{r\sin\vartheta}\frac{\partial U}{\partial \varphi}$
div$\vec{\mathbf{V}}$	$\frac{\partial V_x}{\partial x} + \frac{\partial V_y}{\partial y} + \frac{\partial V_z}{\partial z}$	$\frac{1}{\rho}\frac{\partial}{\partial \rho}(\rho V_\rho) + \frac{1}{\rho}\frac{\partial V_\varphi}{\partial \varphi} + \frac{\partial V_z}{\partial z}$	$\frac{1}{r^2}\frac{\partial}{\partial r}(r^2 V_r) + \frac{1}{r\sin\vartheta}\frac{\partial}{\partial \vartheta}(V_\vartheta \sin\vartheta)$ $+\frac{1}{r\sin\vartheta}\frac{\partial V_\varphi}{\partial \varphi}$
rot$\vec{\mathbf{V}}$	$\vec{\mathbf{e}}_x \left(\frac{\partial V_z}{\partial y} - \frac{\partial V_y}{\partial z}\right)$ $+\vec{\mathbf{e}}_y \left(\frac{\partial V_x}{\partial z} - \frac{\partial V_z}{\partial x}\right)$ $+\vec{\mathbf{e}}_z \left(\frac{\partial V_y}{\partial x} - \frac{\partial V_x}{\partial y}\right)$	$\vec{\mathbf{e}}_\rho \left(\frac{1}{\rho}\frac{\partial V_z}{\partial \varphi} - \frac{\partial V_\varphi}{\partial z}\right)$ $+\vec{\mathbf{e}}_\varphi \left(\frac{\partial V_\rho}{\partial z} - \frac{\partial V_z}{\partial \rho}\right)$ $+\vec{\mathbf{e}}_z \left(\frac{1}{\rho}\frac{\partial}{\partial \rho}(\rho V_\varphi) - \frac{1}{\rho}\frac{\partial V_\rho}{\partial \varphi}\right)$	$\vec{\mathbf{e}}_r \frac{1}{r\sin\vartheta}\left[\frac{\partial}{\partial \vartheta}(V_\varphi \sin\vartheta) - \frac{\partial V_\vartheta}{\partial \varphi}\right]$ $+\vec{\mathbf{e}}_\vartheta \frac{1}{r}\left[\frac{1}{\sin\vartheta}\frac{\partial V_r}{\partial \varphi} - \frac{\partial}{\partial r}(rV_\varphi)\right]$ $+\vec{\mathbf{e}}_\varphi \frac{1}{r}\left[\frac{\partial}{\partial r}(rV_\vartheta) - \frac{\partial V_r}{\partial \vartheta}\right]$
ΔU	$\frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} + \frac{\partial^2 U}{\partial z^2}$	$\frac{1}{\rho}\frac{\partial}{\partial \rho}\left(\rho\frac{\partial U}{\partial \rho}\right) + \frac{1}{\rho^2}\frac{\partial^2 U}{\partial \varphi^2}$ $+\frac{\partial^2 U}{\partial z^2}$	$\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial U}{\partial r}\right)$ $+\frac{1}{r^2\sin\vartheta}\frac{\partial}{\partial \vartheta}\left(\sin\vartheta\frac{\partial U}{\partial \vartheta}\right)$ $+\frac{1}{r^2\sin^2\vartheta}\frac{\partial^2 U}{\partial \varphi^2}$

13.2.7.3 Fundamental Relations and Results (see Table 13.3)

Table 13.3 Fundamental relations for spatial differential operators

Operator	Symbol	Relation	Argument	Result	Meaning
Gradient	grad U	∇U	scalar	vector	maximal increase
Vector gradient	grad $\vec{\mathbf{V}}$	$\nabla\vec{\mathbf{V}}$	vector	tensor second order	
Divergence	div $\vec{\mathbf{V}}$	$\nabla\cdot\vec{\mathbf{V}}$	vector	scalar	source, sink
Rotation	rot $\vec{\mathbf{V}}$	$\nabla\times\vec{\mathbf{V}}$	vector	vector	curl
Laplace operator	ΔU	$(\nabla\cdot\nabla)U$	scalar	scalar	potential field source
Laplace operator	$\Delta\vec{\mathbf{V}}$	$(\nabla\cdot\nabla)\vec{\mathbf{V}}$	vector	vector	

13.3 Integration in Vector Fields

Integration in vector fields is usually performed in Cartesian, cylindrical or in spherical coordinate systems. Usually one integrates along curves, surfaces, or volumes. The line, surface, and volume elements needed for these calculations are collected in **Table 13.4**.

Table 13.4 Line, surface, and volume elements in Cartesian, cylindrical, and spherical coordinates

	Cartesian coordinates	Cylindrical coordinates	Spherical coordinates
$d\vec{\mathbf{r}}$	$\vec{\mathbf{e}}_x dx + \vec{\mathbf{e}}_y dy + \vec{\mathbf{e}}_z dz$	$\vec{\mathbf{e}}_\rho d\rho + \vec{\mathbf{e}}_\varphi \rho d\varphi + \vec{\mathbf{e}}_z dz$	$\vec{\mathbf{e}}_r dr + \vec{\mathbf{e}}_\vartheta r d\vartheta + \vec{\mathbf{e}}_\varphi r\sin\vartheta d\varphi$
$d\vec{\mathbf{S}}$	$\vec{\mathbf{e}}_x dydz + \vec{\mathbf{e}}_y dxdz + \vec{\mathbf{e}}_z dxdy$	$\vec{\mathbf{e}}_\rho \rho d\varphi dz + \vec{\mathbf{e}}_\varphi d\rho dz + \vec{\mathbf{e}}_z \rho d\rho d\varphi$	$\vec{\mathbf{e}}_r r^2\sin\vartheta d\vartheta d\varphi$ $+\vec{\mathbf{e}}_\vartheta r\sin\vartheta dr d\varphi$ $+\vec{\mathbf{e}}_\varphi r dr d\vartheta d\varphi$
dv^*	$dxdydz$	$\rho d\rho d\varphi dz$	$r^2\sin\vartheta dr d\vartheta d\varphi$
	$\vec{\mathbf{e}}_x = \vec{\mathbf{e}}_y \times \vec{\mathbf{e}}_z$ $\vec{\mathbf{e}}_y = \vec{\mathbf{e}}_z \times \vec{\mathbf{e}}_x$ $\vec{\mathbf{e}}_z = \vec{\mathbf{e}}_x \times \vec{\mathbf{e}}_y$	$\vec{\mathbf{e}}_\rho = \vec{\mathbf{e}}_\varphi \times \vec{\mathbf{e}}_z$ $\vec{\mathbf{e}}_\varphi = \vec{\mathbf{e}}_z \times \vec{\mathbf{e}}_\rho$ $\vec{\mathbf{e}}_z = \vec{\mathbf{e}}_\rho \times \vec{\mathbf{e}}_\varphi$	$\vec{\mathbf{e}}_r = \vec{\mathbf{e}}_\vartheta \times \vec{\mathbf{e}}_\varphi$ $\vec{\mathbf{e}}_\vartheta = \vec{\mathbf{e}}_\varphi \times \vec{\mathbf{e}}_r$ $\vec{\mathbf{e}}_\varphi = \vec{\mathbf{e}}_r \times \vec{\mathbf{e}}_\vartheta$
	$\vec{\mathbf{e}}_i\cdot\vec{\mathbf{e}}_j = \begin{cases}0 & i\neq j\\ 1 & i=j\end{cases}$	$\vec{\mathbf{e}}_i\cdot\vec{\mathbf{e}}_j = \begin{cases}0 & i\neq j\\ 1 & i=j\end{cases}$	$\vec{\mathbf{e}}_i\cdot\vec{\mathbf{e}}_j = \begin{cases}0 & i\neq j\\ 1 & i=j\end{cases}$
	The indices i and j take the place of x, y, z or ρ, φ, z or r, ϑ, φ.		
*	The volume is denoted here by v to avoid confusion with the absolute value of the vector function $\lvert\vec{\mathbf{V}}\rvert = V$.		

13.3.1 Line Integral and Potential in Vector Fields

13.3.1.1 Line Integral in Vector Fields

1. Definition The scalar-valued curvilinear integral or line integral of a vector function $\vec{\mathbf{V}}(\vec{\mathbf{r}})$ along a rectifiable curve $\widehat{AB}$ **(Fig. 13.13)** is the scalar value

$$P = \int\limits_{\widehat{AB}} \vec{\mathbf{V}}(\vec{\mathbf{r}})\cdot d\vec{\mathbf{r}}. \tag{13.99a}$$

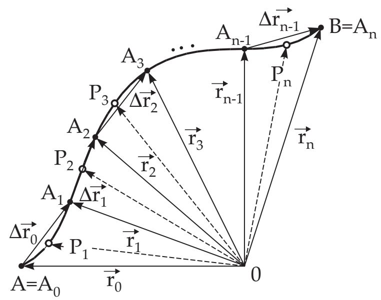


Figure 13.13

2. Evaluation of this Integral in Five Steps

a) Dividing the path $\overset{\frown}{AB}$ **(Fig. 13.13)** by division points $A_1(\vec{\mathbf{r}}_1),\ A_2(\vec{\mathbf{r}}_2), \ldots, A_{n-1}(\vec{\mathbf{r}}_{n-1})$ $(A = A_0,\ B = A_n)$ into n small arcs which are approximated by the vectors $\vec{\mathbf{r}}_i - \vec{\mathbf{r}}_{i-1} = \Delta\vec{\mathbf{r}}_{i-1}$.

b) Choosing arbitrarily the points P_i with position vectors $\vec{\xi}_i$ lying inside or at the boundary of each small arc.

c) Calculating the dot product of the value of the function $\vec{\mathbf{V}}(\vec{\xi}_i)$ at these chosen points with the corresponding $\Delta\vec{\mathbf{r}}_{i-1}$.

d) Taking the sum of all the n products.

e) Calculating the limit of the sums got in this way $\sum\limits_{i=1}^{n} \tilde{\vec{\mathbf{V}}}(\vec{\xi}_i) \cdot \Delta\vec{\mathbf{r}}_{i-1}$ for $|\Delta\vec{\mathbf{r}}_{i-1}| \to 0$, while $n \to \infty$ obviously.

If this limit exists independently of the choice of the points A_i and P_i, then it is called the line integral

$$\int\limits_{\overset{\frown}{AB}} \vec{\mathbf{V}} \cdot d\vec{\mathbf{r}} = \lim_{\substack{|\Delta\vec{\mathbf{r}}_{i-1}| \to 0 \\ n \to \infty}} \sum_{i=1}^{n} \tilde{\vec{\mathbf{V}}}(\vec{\xi}_i) \cdot \Delta\vec{\mathbf{r}}_{i-1}. \tag{13.99b}$$

A sufficient condition for the existence of the line integral (13.99a,b) is that the vector function $\vec{\mathbf{V}}(\vec{\mathbf{r}})$ and the curve $\overset{\frown}{AB}$ are continuous and the curve has a tangent varying continuously. A vector function $\vec{\mathbf{V}}(\vec{\mathbf{r}})$ is continuous if its components, the three scalar functions, are continuous.

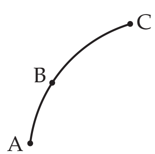


Figure 13.14

13.3.1.2 Interpretation of the Line Integral in Mechanics

If $\vec{\mathbf{V}}(\vec{\mathbf{r}})$ is a field of force, i.e., $\vec{\mathbf{V}}(\vec{\mathbf{r}}) = \vec{\mathbf{F}}(\vec{\mathbf{r}})$, then the line integral (13.99a) represents the work done by $\vec{\mathbf{F}}$ while a particle m moves along the path $\overset{\frown}{AB}$ **(Fig. 13.13,13.14)**.

13.3.1.3 Properties of the Line Integral

$$\int\limits_{\overset{\frown}{ABC}} \vec{\mathbf{V}}(\vec{\mathbf{r}}) \cdot d\vec{\mathbf{r}} = \int\limits_{\overset{\frown}{AB}} \vec{\mathbf{V}}(\vec{\mathbf{r}}) \cdot d\vec{\mathbf{r}} + \int\limits_{\overset{\frown}{BC}} \vec{\mathbf{V}}(\vec{\mathbf{r}}) \cdot d\vec{\mathbf{r}} \qquad \textbf{(Fig. 13.14)}. \tag{13.100}$$

$$\int\limits_{\overset{\frown}{AB}} \vec{\mathbf{V}}(\vec{\mathbf{r}}) \cdot d\vec{\mathbf{r}} = -\int\limits_{\overset{\frown}{BA}} \vec{\mathbf{V}}(\vec{\mathbf{r}}) \cdot d\vec{\mathbf{r}}. \tag{13.101}$$

$$\int\limits_{\overset{\frown}{AB}} \left[\vec{\mathbf{V}}(\vec{\mathbf{r}}) + \vec{\mathbf{W}}(\vec{\mathbf{r}})\right] \cdot d\vec{\mathbf{r}} = \int\limits_{\overset{\frown}{AB}} \vec{\mathbf{V}}(\vec{\mathbf{r}}) \cdot d\vec{\mathbf{r}} + \int\limits_{\overset{\frown}{AB}} \vec{\mathbf{W}}(\vec{\mathbf{r}}) \cdot d\vec{\mathbf{r}}. \tag{13.102}$$

$$\int\limits_{\overset{\frown}{AB}} c\vec{\mathbf{V}}(\vec{\mathbf{r}}) \cdot d\vec{\mathbf{r}} = c \int\limits_{\overset{\frown}{AB}} \vec{\mathbf{V}}(\vec{\mathbf{r}}) \cdot d\vec{\mathbf{r}} \qquad \text{(c const)}. \tag{13.103}$$

13.3.1.4 Line Integral in Cartesian Coordinates

In Cartesian coordinates the following formula holds:

$$\int_{\widehat{AB}} \vec{\mathbf{V}}(\vec{\mathbf{r}}) \cdot d\vec{\mathbf{r}} = \int_{\widehat{AB}} (V_x\,dx + V_y\,dy + V_z\,dz)\,. \tag{13.104}$$

13.3.1.5 Integral Along a Closed Curve in a Vector Field

A line integral is called a *contour integral* if the path of integration is a closed curve. If the scalar value of the integral is denoted by P and the closed curve is denoted by C, then the following notation is used:

$$P = \oint_{(C)} \vec{\mathbf{V}}(\vec{\mathbf{r}}) \cdot d\vec{\mathbf{r}}. \tag{13.105}$$

13.3.1.6 Conservative Field or Potential Field

1. Definition

If the value P of the line integral (13.99a) in a vector field depends only on the initial point A and the endpoint B, and is independent of the path between them, then this field is called a *conservative field* or a *potential field*.

The value of the contour integral in a conservative field is always equal to zero:

$$\int_{(C)} \vec{\mathbf{V}}(\vec{\mathbf{r}}) \cdot d\vec{\mathbf{r}} = 0. \tag{13.106}$$

A conservative field is always irrotational:

$$\operatorname{rot} \vec{\mathbf{V}} = \vec{\mathbf{0}}, \tag{13.107}$$

and conversely, this equality is a sufficient condition for a vector field to be conservative. Of course, it is to be supposed that the partial derivatives of the field function $\vec{\mathbf{V}}$ are continuous with respect to the corresponding coordinates, and the domain of $\vec{\mathbf{V}}$ is simply connected. This condition, also called the *integrability condition* (see 8.3.4.2, p. 521), has the following form in Cartesian coordinates

$$\frac{\partial V_x}{\partial y} = \frac{\partial V_y}{\partial x}\,, \quad \frac{\partial V_y}{\partial z} = \frac{\partial V_z}{\partial y}\,, \quad \frac{\partial V_z}{\partial x} = \frac{\partial V_x}{\partial z}\,. \tag{13.108}$$

2. Potential of a Conservative Field,

or its potential function or briefly its potential is the scalar function

$$U(\vec{\mathbf{r}}) = \int_{\vec{\mathbf{r}}_0}^{\vec{\mathbf{r}}} \vec{\mathbf{V}}(\vec{\mathbf{r}}) \cdot d\vec{\mathbf{r}}. \tag{13.109a}$$

In a conservative field it is calculated with a fixed initial point $A(\vec{\mathbf{r}}_0)$ and a variable endpoint $B(\vec{\mathbf{r}})$ as the line integral

$$U(\vec{\mathbf{r}}) = \int_{\widehat{AB}} \vec{\mathbf{V}}(\vec{\mathbf{r}}) \cdot d\vec{\mathbf{r}}. \tag{13.109b}$$

Remark: In physics, the potential $U^*(\vec{\mathbf{r}})$ of a function $\vec{\mathbf{V}}(\vec{\mathbf{r}})$ at the point $\vec{\mathbf{r}}$ is often considered with the opposite sign:

$$U^*(\vec{\mathbf{r}}) = -\int_{\vec{\mathbf{r}}_0}^{\vec{\mathbf{r}}} \vec{\mathbf{V}}(\vec{\mathbf{r}}) \cdot d\vec{\mathbf{r}} = -U(\vec{\mathbf{r}}). \tag{13.110}$$

3. Relations between Gradient, Line Integral, and Potential

If the relation $\vec{\mathbf{V}}(\vec{\mathbf{r}}) = \operatorname{grad} U(\vec{\mathbf{r}})$ holds, then $U(\vec{\mathbf{r}})$ is the potential of the field $\vec{\mathbf{V}}(\vec{\mathbf{r}})$, and conversely, $\vec{\mathbf{V}}(\vec{\mathbf{r}})$ is a conservative or potential field. In physics often the negative sign is used corresponding to (13.110).

4. Calculation of the Potential in a Conservative Field

If the function $\vec{\mathbf{V}}(\vec{\mathbf{r}})$ is given in Cartesian coordinates $\vec{\mathbf{V}} = V_x\vec{\mathbf{i}} + V_y\vec{\mathbf{j}} + V_z\vec{\mathbf{k}}$, then for the total differential of its potential function U

$$dU = V_x\,dx + V_y\,dy + V_z\,dz \tag{13.111a}$$

holds. Here, the coefficients V_x, V_y, V_z must fulfill the integrability condition (13.108). The determination of U follows from the equation system

$$\frac{\partial U}{\partial x} = V_x, \quad \frac{\partial U}{\partial y} = V_y, \quad \frac{\partial U}{\partial z} = V_z. \tag{13.111b}$$

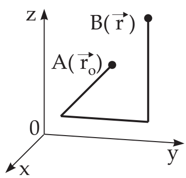


Figure 13.15

In practice, the calculation of the potential can be done by performing the integration along three straight line segments parallel to the coordinate axes and connected to each other **(Fig. 13.15)**:

$$U = \int_{\vec{\mathbf{r}}_0}^{\vec{\mathbf{r}}} \vec{\mathbf{V}} \cdot d\vec{\mathbf{r}} = U(x_0, y_0, z_0) + \int_{x_0}^{x} V_x(x, y_0, z_0)\,dx + \int_{y_0}^{y} V_y(x, y, z_0)\,dy + \int_{z_0}^{z} V_z(x, y, z)\,dz\,. \tag{13.112}$$

13.3.2 Surface Integrals

13.3.2.1 Vector of a Plane Sheet

The vector representation of the surface integral of general type (see 8.3.4.2, p. 537) requires to assign a vector $\vec{\mathbf{S}}$ to a plane surface region S, which is perpendicular to this region and its absolute value is equal to the area of S. **Fig. 13.16a** shows the case of a plane sheet. The positive direction in S is given by defining the positive sense along a closed curve C according to the *right-hand law* (also called *right-screw rule*): Looking from the initial point of the vector into the direction of its final point, then the *positive sense* is the clockwise direction. By this choice of orientation of the boundary curve one fixes the exterior side of this surface region, i.e., the side on which the vector lies. This definition works in the case of any surface region bounded by a closed curve **(Fig. 13.16b,c)**.

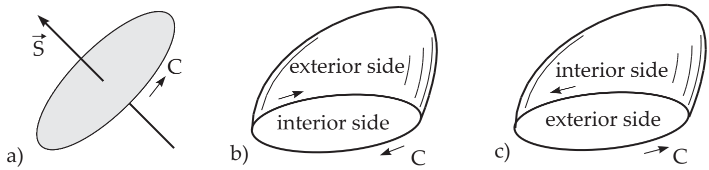


Figure 13.16

13.3.2.2 Evaluation of the Surface Integral

The evaluation of a surface integral in scalar or vector fields is independent of whether the surface S is bounded by a closed curve or is itself a closed surface. The evaluation is performed in five steps:

a) Dividing the surface region S on the exterior side defined by the orientation of the boundary curve **(Fig. 13.17)** into n arbitrary elementary surfaces ΔS_i so that each of these surface elements can be approximated by a plane surface element. Assigning the vector $\Delta\vec{\mathbf{S}}_{\mathbf{i}}$ to every surface element ΔS_i as given in (13.33a). In the case of a closed surface, the positive direction is defined so that the exterior

side is where $\Delta\vec{\mathbf{S}}_\mathbf{i}$ should start.

b) Choosing an arbitrary point P_i with the position vector $\vec{\mathbf{r}}_\mathbf{i}$ inside or on the boundary of each surface element.

c) Producing the products $U(\vec{\mathbf{r}}_\mathbf{i})\,\Delta\vec{\mathbf{S}}_\mathbf{i}$ in the case of a scalar field and the product $\vec{\mathbf{V}}(\vec{\mathbf{r}}_\mathbf{i})\cdot\Delta\vec{\mathbf{S}}_\mathbf{i}$ or $\vec{\mathbf{V}}(\vec{\mathbf{r}}_\mathbf{i})\times\Delta\vec{\mathbf{S}}_\mathbf{i}$ in the case of a vector field.

d) Taking the sum of all these products.

e) Evaluating the limit while the diameters of ΔS_i tend to zero, i.e., $|\Delta\vec{\mathbf{S}}_\mathbf{i}| \to 0$ for $n \to \infty$. So, the surface elements tend to zero in the sense given in 8.4.1, **1.**, p. 524, for double integrals.
If this limit exists independently of the partition and of the choice of the points $\vec{\mathbf{r}}_\mathbf{i}$, then one calls it the surface integral of $\vec{\mathbf{V}}$ on the given surface.

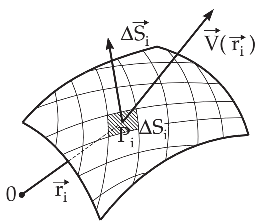


Figure 13.17

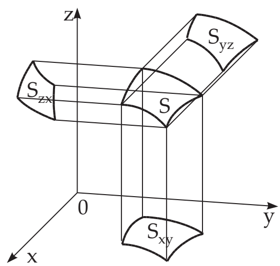


Figure 13.18

13.3.2.3 Surface Integrals and Flow of Fields

1. Vector Flow of a Scalar Field

$$\vec{\mathbf{P}} = \lim_{\substack{|\Delta\vec{\mathbf{S}}_i|\to 0\\ n\to\infty}} \sum_{i=1}^{n} U(\vec{\mathbf{r}}_i)\,\Delta\vec{\mathbf{S}}_i = \int_{(S)} U(\vec{\mathbf{r}})\,d\vec{\mathbf{S}}. \tag{13.113}$$

2. Scalar Flow of a Vector Field

$$Q = \lim_{\substack{|\Delta\vec{\mathbf{S}}_i|\to 0\\ n\to\infty}} \sum_{i=1}^{n} \vec{\mathbf{V}}(\vec{\mathbf{r}}_i)\cdot\Delta\vec{\mathbf{S}}_i = \int_{(S)} \vec{\mathbf{V}}(\vec{\mathbf{r}})\cdot d\vec{\mathbf{S}}. \tag{13.114}$$

3. Vector Flow of a Vector Field

$$\vec{\mathbf{R}} = \lim_{\substack{|\Delta\vec{\mathbf{S}}_i|\to 0\\ n\to\infty}} \sum_{i=1}^{n} \vec{\mathbf{V}}(\vec{\mathbf{r}}_i)\times\Delta\vec{\mathbf{S}}_i = \int_{(S)} \vec{\mathbf{V}}(\vec{\mathbf{r}})\times d\vec{\mathbf{S}}. \tag{13.115}$$

13.3.2.4 Surface Integral in Cartesian Coordinates as Surface Integrals of Second Type

$$\int_{(S)} U\,d\vec{\mathbf{S}} = \iint_{(S_{yz})} U\,dy\,dz\,\vec{\mathbf{i}} + \iint_{(S_{zx})} U\,dz\,dx\,\vec{\mathbf{j}} + \iint_{(S_{xy})} U\,dx\,dy\,\vec{\mathbf{k}}. \tag{13.116}$$

$$\int_{(S)} \vec{\mathbf{V}}\cdot d\vec{\mathbf{S}} = \iint_{(S_{yz})} V_x\,dy\,dz + \iint_{(S_{zx})} V_y\,dz\,dx + \iint_{(S_{xy})} V_z\,dx\,dy. \tag{13.117}$$

$$\int_{(S)} \vec{\mathbf{V}} \times d\vec{\mathbf{S}} = \iint_{(S_{yz})} (V_z\vec{\mathbf{j}} - V_y\vec{\mathbf{k}})\, dy\, dz + \iint_{(S_{zx})} (V_x\vec{\mathbf{k}} - V_z\vec{\mathbf{i}})\, dz\, dx + \iint_{(S_{xy})} (V_y\vec{\mathbf{i}} - V_x\vec{\mathbf{j}})\, dx\, dy. \quad (13.118)$$

The existence theorems for these integrals can be given similarly to those in 8.5.2, **4.**, p. 537.
In the formulas above, each of the integrals is taken over the projection S on the corresponding coordinate plane **(Fig. 13.18)**, where one of the variables x, y or z should be expressed by the others from the equation of S.

Remark: Integrals over a closed surface are denoted by

$$\oint_{(S)} U\, d\vec{\mathbf{S}} = \oiint_{(S)} U\, d\vec{\mathbf{S}}, \qquad \oint_{(S)} \vec{\mathbf{V}} \cdot d\vec{\mathbf{S}} = \oiint_{(S)} \vec{\mathbf{V}} \cdot d\vec{\mathbf{S}}, \qquad \oint_{(S)} \vec{\mathbf{V}} \times d\vec{\mathbf{S}} = \oiint_{(S)} \vec{\mathbf{V}} \times d\vec{\mathbf{S}}. \quad (13.119)$$

■ **A:** Calculate the integral $\vec{\mathbf{P}} = \int_{(S)} xyz\, d\vec{\mathbf{S}}$, where the surface is the plane region $x+y+z=1$ bounded by the coordinate planes. The upward side is the positive side:
$\vec{\mathbf{P}} = \iint_{(S_{yz})} (1-y-z)yz\, dy\, dz\, \vec{\mathbf{i}} + \iint_{(S_{zx})} (1-x-z)xz\, dz\, dx\, \vec{\mathbf{j}} + \iint_{(S_{xy})} (1-x-y)xy\, dx\, dy\, \vec{\mathbf{k}}$;
$\iint_{(S_{yz})} (1-y-z)yz\, dy\, dz = \int_0^1 \int_0^{1-z} (1-y-z)yz\, dy\, dz = \frac{1}{120}$. We get the two further integrals analogously. The result is: $\vec{\mathbf{P}} = \frac{1}{120}(\vec{\mathbf{i}} + \vec{\mathbf{j}} + \vec{\mathbf{k}})$.

■ **B:** Calculate the integral $Q = \int_{(S)} \vec{\mathbf{r}} \cdot d\vec{\mathbf{S}} = \iint_{(S_{yz})} x\, dy\, dz + \iint_{(S_{zx})} y\, dz\, dx + \iint_{(S_{xy})} z\, dx\, dy$ over the same plane region as in **A**: $\iint_{(S_{yz})} x\, dy\, dz = \int_0^1 \int_0^{1-z} (1-y-z)\, dy\, dz = \frac{1}{6}$. Both other integrals are calculated similarly. The result is: $Q = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}$.

■ **C:** Calculate the integral $\vec{\mathbf{R}} = \int_{(S)} \vec{\mathbf{r}} \times d\vec{\mathbf{S}} = \int_{(S)} (x\vec{\mathbf{i}} + y\vec{\mathbf{j}} + z\vec{\mathbf{k}}) \times (dy\, dz\, \vec{\mathbf{i}} + dz\, dx\, \vec{\mathbf{j}} + dx\, dy\, \vec{\mathbf{k}})$, where the surface region is the same as in **A**: Performing the computations gives $\vec{\mathbf{R}} = \vec{\mathbf{0}}$.

13.3.3 Integral Theorems

13.3.3.1 Integral Theorem and Integral Formula of Gauss

1. Integral Theorem of Gauss or the Divergence Theorem

The *integral theorem of Gauss* gives the relation between a volume integral of the divergence of $\vec{\mathbf{V}}$ over a volume v, and a surface integral over the surface S surrounding this volume. The orientation of the surface (see 8.5.2.1, p. 535) is defined so that the exterior side is the positive one. The vector function $\vec{\mathbf{V}}$ should be continuous, their first partial derivatives should exist and be continuous. The integral theorem of Gauss reads as follows:

$$\oiint_{(S)} \vec{\mathbf{V}} \cdot d\vec{\mathbf{S}} = \iiint_{(v)} \operatorname{div} \vec{\mathbf{V}}\, dv, \quad (13.120a)$$

i.e., the scalar flow of the field $\vec{\mathbf{V}}$ through a closed surface S is equal to the integral of divergence of $\vec{\mathbf{V}}$ over the volume v bounded by S. In Cartesian coordinates one gets:

$$\oiint\limits_{(S)}(V_x\,dy\,dz + V_y\,dz\,dx + V_z\,dx\,dy) = \iiint\limits_{(v)}\left(\frac{\partial V_x}{\partial x} + \frac{\partial V_y}{\partial y} + \frac{\partial V_z}{\partial z}\right)dx\,dy\,dz\,. \qquad (13.120b)$$

2. Integral Formula of Gauss

In the planar case, the integral theorem of Gauss restricted to the x, y plane becomes the *integral formula of Gauss*. It represents the correspondence between a line integral and the corresponding surface integral. The integral formula of Gauss reads as follows:

$$\iint\limits_{(B)}\left[\frac{\partial Q(x,y)}{\partial x} - \frac{\partial P(x,y)}{\partial y}\right]dx\,dy = \oint\limits_{(C)}[P(x,y)\,dx + Q(x,y)\,dy]\,. \qquad (13.121)$$

B denotes a plane region which is bounded by C. P and Q are continuous functions with continuous first partial derivatives.

3. Sector Formula

The *sector formula* is an important special case of the Gauss integral formula to calculate the area of plane regions. For $Q = x$, $P = -y$ it follows that

$$F = \iint\limits_{(B)} dx\,dy = \frac{1}{2}\oint\limits_{(C)}[x\,dy - y\,dx]. \qquad (13.122)$$

13.3.3.2 Integral Theorem of Stokes

The *integral theorem of Stokes* gives the relation between a surface integral over an oriented surface region S, in which the vector field $\vec{\mathbf{V}}$ is defined, and the integral along the closed boundary curve C of the surface S. The sense of the curve C is chosen so that the sense of traverse forms a *right-screw* with the surface normal (see 13.3.2.1, p. 722). The vector function $\vec{\mathbf{V}}$ should be continuous and it should have continuous first partial derivatives. The integral theorem of Stokes reads as follows:

$$\iint\limits_{(S)}\operatorname{rot}\vec{\mathbf{V}}\cdot d\vec{\mathbf{S}} = \oint\limits_{(C)}\vec{\mathbf{V}}\cdot d\vec{\mathbf{r}}, \qquad (13.123a)$$

i.e., the vector flow of the rotation through a surface S bounded by the closed curve C is equal to the contour integral of the vector field $\vec{\mathbf{V}}$ along the curve C.
In Cartesian coordinates

$$\iint\limits_{(S)}\left[\left(\frac{\partial V_z}{\partial y} - \frac{\partial V_y}{\partial z}\right)dy\,dz + \left(\frac{\partial V_x}{\partial z} - \frac{\partial V_z}{\partial x}\right)dz\,dx + \left(\frac{\partial V_y}{\partial x} - \frac{\partial V_x}{\partial y}\right)dx\,dy\right]$$

$$= \oint\limits_{(C)}(V_x\,dx + V_y\,dy + V_z\,dz) \qquad (13.123b)$$

holds. In the planar case, the integral theorem of Stokes, just as that of Gauss, becomes into the integral formula (13.121) of Gauss.

13.3.3.3 Integral Theorems of Green

The Green integral theorems give relations between volume and surface integrals. They are the applications of the Gauss theorem for the function $\vec{\mathbf{V}} = U_1 \operatorname{grad} U_2$, where U_1 and U_2 are scalar field functions and v is the volume surrounded by the surface S. The following theorems hold:

1. $$\iiint\limits_{(v)}(U_1\Delta U_2 + \operatorname{grad} U_2\cdot\operatorname{grad} U_1)\,dv = \oiint\limits_{(S)}U_1\operatorname{grad} U_2\cdot d\vec{\mathbf{S}}, \qquad (13.124)$$

2. $$\iiint\limits_{(v)} (U_1 \Delta U_2 - U_2 \Delta U_1)\, dv = \oiint\limits_{(S)} (U_1 \operatorname{grad} U_2 - U_2 \operatorname{grad} U_1) \cdot d\vec{\mathbf{S}}. \tag{13.125}$$

In particular for $U_1 = 1, U_2 = U$

3. $$\iiint\limits_{(v)} \Delta U\, dv = \oiint\limits_{(S)} \operatorname{grad} U \cdot d\vec{\mathbf{S}} \tag{13.126}$$

holds. In Cartesian coordinates the third Green theorem has the following form (compare (13.120b)):

$$\iiint\limits_{(v)} \left(\frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} + \frac{\partial^2 U}{\partial z^2} \right) dv = \oiint\limits_{(S)} \left(\frac{\partial U}{\partial x}\, dy\, dz + \frac{\partial U}{\partial y}\, dz\, dx + \frac{\partial U}{\partial z}\, dx\, dy \right). \tag{13.127}$$

■ **A:** Calculating the line integral $I = \oint\limits_{(C)} (x^2 y^3\, dx + dy + z\, dz)$ with a circle C as the intersection curve of the cylinder $x^2 + y^2 = a^2$ and the plane $z = 0$. With the Stokes theorem (13.123a) one gets:

$$I = \oint\limits_{(C)} \vec{\mathbf{V}} \cdot d\vec{\mathbf{r}} = \iint\limits_{(S)} \operatorname{rot} \vec{\mathbf{V}} \cdot d\vec{\mathbf{S}} = -\iint\limits_{(S^*)} 3x^2 y^2\, dx\, dy = -3 \int\limits_{\varphi=0}^{2\pi} \int\limits_{r=0}^{a} r^5 \cos^2\varphi \sin^2\varphi\, dr\, d\varphi = -\frac{a^6}{8}\pi \quad \text{with}$$

$\operatorname{rot} \vec{\mathbf{V}} = -3x^2 y^2 \vec{\mathbf{k}}$, $d\vec{\mathbf{S}} = \vec{\mathbf{k}}\, dx\, dy$ and the circle S^*: $x^2 + y^2 \le a^2$.

■ **B:** Determine the flux $I = \oint\limits_{(S)} \vec{\mathbf{V}} \cdot d\vec{\mathbf{S}}$ in the drift space $\vec{\mathbf{V}} = x^3 \vec{\mathbf{i}} + y^3 \vec{\mathbf{j}} + z^3 \vec{\mathbf{k}}$ through the surface S of the sphere $x^2 + y^2 + z^2 = a^2$. The theorem of Gauss yields:

$$I = \oint\limits_{(S)} \vec{\mathbf{V}} \cdot d\vec{\mathbf{S}} = \iiint\limits_{(v)} \operatorname{div} \vec{\mathbf{V}}\, dv = 3 \iiint\limits_{(v)} (x^2 + y^2 + z^2)\, dx\, dy\, dz = 3 \int\limits_{\varphi=0}^{2\pi} \int\limits_{\vartheta=0}^{\pi} \int\limits_{r=0}^{a} r^4 \sin\vartheta\, dr\, d\vartheta\, d\varphi = \frac{12}{5} a^5 \pi.$$

■ **C:** Heat conduction equation: The change in time of the heat Q of a space region v containing no heat source is given by $\dfrac{dQ}{dt} = \iiint\limits_{(v)} c\varrho \dfrac{\partial T}{\partial t}\, dv$ (specific heat-capacity c, density ϱ, temperature T), while the corresponding time-dependent change of the heat flow through the surface S of v is given by $\dfrac{dQ}{dt} = \oiint\limits_{(S)} \lambda \operatorname{grad} T \cdot d\vec{\mathbf{S}}$ (thermal conductivity λ). Applying the theorem of Gauss for the surface integral (13.120a) one gets from $\iiint\limits_{(v)} \left[c\varrho \dfrac{\partial T}{\partial t} - \operatorname{div}(\lambda \operatorname{grad} T) \right] dv = 0$ the heat conduction equation $c\lambda \dfrac{\partial T}{\partial t} = \operatorname{div}(\lambda \operatorname{grad} T)$, which has the form $\dfrac{\partial T}{\partial t} = a^2 \Delta T$ in the case of a homogeneous solid (c, ϱ, λ constants).

13.4 Evaluation of Fields

13.4.1 Pure Source Fields

A field $\vec{\mathbf{V}}_1$ is called a *pure source field* or an *irrotational source field* when its rotation is equal to zero everywhere. If the *divergence* is $q(\vec{\mathbf{r}})$, then

$$\operatorname{div} \vec{\mathbf{V}}_1 = q(\vec{\mathbf{r}}), \qquad \operatorname{rot} \vec{\mathbf{V}}_1 \equiv \vec{\mathbf{0}} \tag{13.128}$$

holds. In this case, the field has a potential U, which is defined at every point P by the *Poisson differential equation* (see 13.5.2, p. 729)

$$\vec{\mathbf{V}}_1 = \operatorname{grad} U, \qquad \operatorname{div}\operatorname{grad} U = \Delta U = q(\vec{\mathbf{r}}), \tag{13.129a}$$

where $\vec{\mathbf{r}}$ is the position vector of P. (In physics most often $\vec{\mathbf{V}}_1 = -\operatorname{grad} U$ is used.) The evaluation of U comes from

$$U(\vec{\mathbf{r}}) = -\frac{1}{4\pi}\iiint \frac{\operatorname{div}\vec{\mathbf{V}}(\vec{\mathbf{r}}^*)\,dv(\vec{\mathbf{r}}^*)}{|\vec{\mathbf{r}} - \vec{\mathbf{r}}^*|}. \tag{13.129b}$$

The integration is taken over the whole of space **(Fig. 13.19)**. The divergence of $\vec{\mathbf{V}}$ must be differentiable and be decreasing sufficiently rapidly for large distances.

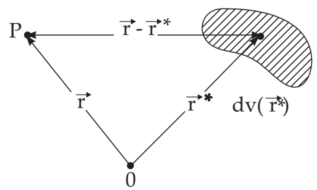


Figure 13.19

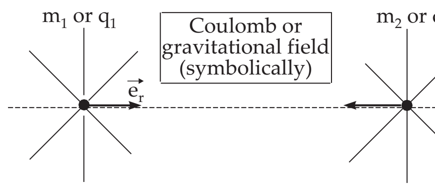


Figure 13.20

13.4.2 Pure Rotation Field or Zero-Divergence Field

A *pure rotation* (or *curl*) *field* or a *solenoidal field* is a vector field $\vec{\mathbf{V}}_2$ whose divergence is equal to zero everywhere; this field is free of sources. With $\vec{\mathbf{w}}(\vec{\mathbf{r}})$ as the *rotation density*

$$\operatorname{div}\vec{\mathbf{V}}_2 \equiv 0, \qquad \operatorname{rot}\vec{\mathbf{V}}_2 = \vec{\mathbf{w}}(\vec{\mathbf{r}}) \tag{13.130a}$$

hold. The rotation density $\vec{\mathbf{w}}(\vec{\mathbf{r}})$ cannot be arbitrary; it must satisfy the equation $\operatorname{div}\vec{\mathbf{w}} = 0$. With the approach

$$\vec{\mathbf{V}}_2(\vec{\mathbf{r}}) = \operatorname{rot}\vec{\mathbf{A}}(\vec{\mathbf{r}}), \qquad \operatorname{div}\vec{\mathbf{A}} = 0, \quad \text{i.e.,} \quad \operatorname{rot}\operatorname{rot}\vec{\mathbf{A}} = \vec{\mathbf{w}} \tag{13.130b}$$

follows according to (13.97)

$$\operatorname{grad}\operatorname{div}\vec{\mathbf{A}} - \Delta\vec{\mathbf{A}} = \vec{\mathbf{w}}, \quad \text{i.e.,} \quad \Delta\vec{\mathbf{A}} = -\vec{\mathbf{w}}. \tag{13.130c}$$

So, $\vec{\mathbf{A}}(\vec{\mathbf{r}})$ formally satisfies the Poisson differential equation (see (13.135a), p. 729) just as the potential U of an irrotational field $\vec{\mathbf{V}}_1$ and that is why it is called a *vector potential*. For every point P, then

$$\vec{\mathbf{V}}_2 = \operatorname{rot}\vec{\mathbf{A}} \qquad \text{holds with} \quad \vec{\mathbf{A}} = \frac{1}{4\pi}\iiint \frac{\vec{\mathbf{w}}(\vec{\mathbf{r}}^*)}{|\vec{\mathbf{r}} - \vec{\mathbf{r}}^*|}\,dv(\vec{\mathbf{r}}^*)\,. \tag{13.130d}$$

The meaning of $\vec{\mathbf{r}}$ is the same as in (13.129b); the integration is taken over the whole of space.

13.4.3 Vector Fields with Point-Like Sources

13.4.3.1 Coulomb Field of a Point-Like Charge

The *Coulomb field* is an example of an irrotational field, which is also solenoidal, except at the location of the point charge q, the point source **(Fig. 13.20)**. For the Coulomb force

$$\vec{\mathbf{F}}_{\mathrm{C}} = \frac{1}{4\pi\varepsilon_0}\frac{q_1 q_2}{r^2}\,\vec{\mathbf{e}}_r = \frac{q_1}{4\pi\varepsilon_0}q_2\frac{\vec{\mathbf{r}}}{r^3} = eq_2\frac{\vec{\mathbf{r}}}{r^3}\,, \qquad e = \frac{q_1}{4\pi\varepsilon_0} \tag{13.131a}$$

holds. This force affects attractively for electric charges q_1, q_2 with different signs and repulsively for charges with equal signs. ε_0 is the electric constant (see **Table 21.2**, p. 1053), e is the intensity or

source strength of the source. The electric field strength and the electrostatic potential, generated in the space around the charge q_1 and affecting to the charge q_2 are given as

$$\vec{\mathbf{E}}_C = \frac{\vec{\mathbf{F}}_C}{q_2} = \frac{e}{r^3}\vec{\mathbf{r}} = -\operatorname{grad} U\,, \quad U = \frac{e}{r}\,. \tag{13.131b}$$

U denotes the electrostatic potential of the field. The scalar flow in accordance with the theorem of Gauss (see (13.120a), p. 724) is equal to $4\pi e$ or 0, depending on whether the surface S encloses the point source or not:

$$\oint_{(S)} \vec{\mathbf{E}} \cdot d\vec{\mathbf{S}} = \begin{cases} 4\pi e, & \text{if } S \text{ encloses the point source,} \\ 0, & \text{otherwise.} \end{cases} \tag{13.131c}$$

Because of the irrotationality of the electrostatic field

$$\operatorname{rot} \vec{\mathbf{E}}_C \equiv \vec{\mathbf{0}}\,. \tag{13.131d}$$

13.4.3.2 Gravitational Field of a Point Mass

The field of gravity of a point mass or the Newton field is a second example of an irrotational and at the same time solenoidal field, except at the point of the center of mass. For the Newton mass attraction

$$\vec{\mathbf{F}}_N = \gamma \frac{m_1 m_2}{r^2}\,\vec{\mathbf{e}}_r \tag{13.132}$$

holds, where γ is the gravitational constant (see **Table 21.2**, p. 1053). Every relation valid for the Coulomb field is valid analogously also for the Newton field.

13.4.4 Superposition of Fields

13.4.4.1 Discrete Source Distribution

Analogously to superposition of fields in physics, vector fields superpose each other. The *superposition law* is: If the vector fields $\vec{\mathbf{V}}_\nu$ have the potentials U_ν, then the vector field

$$\vec{\mathbf{V}} = \Sigma \vec{\mathbf{V}}_\nu \text{ has the potential } U = \Sigma U_\nu. \tag{13.133a}$$

For n discrete point sources with source strength e_ν $(\nu = 1, 2, \ldots, n)$, whose fields are superposed, the resulting field can be determined by the algebraic sum of the potentials U_ν:

$$\vec{\mathbf{V}}(\vec{\mathbf{r}}) = -\operatorname{grad} \sum_{\nu=1}^{n} U_\nu \qquad \text{with} \qquad U_\nu = \frac{e_\nu}{|\vec{\mathbf{r}} - \vec{\mathbf{r}}_\nu|}. \tag{13.133b}$$

Here, the vector $\vec{\mathbf{r}}$ is again the position vector of the point under consideration, $\vec{\mathbf{r}}_\nu$ are the position vectors of the sources.

If there is an irrotational field $\vec{\mathbf{V}}_1$ and a zero-divergence field $\vec{\mathbf{V}}_2$ together and they are everywhere continuous, then

$$\vec{\mathbf{V}} = \vec{\mathbf{V}}_1 + \vec{\mathbf{V}}_2 = -\frac{1}{4\pi}\left[\operatorname{grad} \iiint \frac{q(\vec{\mathbf{r}}^*)}{|\vec{\mathbf{r}} - \vec{\mathbf{r}}^*|}\, dv(\vec{\mathbf{r}}^*) - \operatorname{rot} \iiint \frac{\vec{\mathbf{w}}(\vec{\mathbf{r}}^*)}{|\vec{\mathbf{r}} - \vec{\mathbf{r}}^*|}\, dv(\vec{\mathbf{r}}^*)\right]. \tag{13.133c}$$

If the vector field is extended to infinity, then the decomposition of $\vec{\mathbf{V}}(\vec{\mathbf{r}})$ is unique if $|\vec{\mathbf{V}}(\vec{\mathbf{r}})|$ decreases sufficient rapidly for $r = |\vec{\mathbf{r}}| \to \infty$. The integration is taken over the whole of space.

13.4.4.2 Continuous Source Distribution

If the sources are distributed continuously along lines, surfaces, or in domains of space, then, instead of the finite source strength e_ν, there are infinitesimals corresponding to the density of the source distributions, and instead of the sums, we have integrals over the domain. In the case of a continuous space distribution of source strength, the divergence is $q(\vec{\mathbf{r}}) = \operatorname{div} \vec{\mathbf{V}}$.

Similar statements are valid for the potential of a field defined by rotation. In the case of a continuous space rotation distribution, the " *rotation density* " is defined by $\vec{\mathbf{w}}(\vec{\mathbf{r}}) = \operatorname{rot} \vec{\mathbf{V}}$.

13.4.4.3 Conclusion

A vector field is determined uniquely by its sources and rotations in space if all these sources and rotations located inside a finite space.

13.5 Differential Equations of Vector Field Theory

13.5.1 Laplace Differential Equation

The problem to determine the potential U of a vector field $\vec{\mathbf{V}}_1 = \operatorname{grad} U$ containing no sources, leads to the equation according to (13.128) with $q(\vec{\mathbf{r}}) = 0$

$$\operatorname{div} \vec{\mathbf{V}}_1 = \operatorname{div} \operatorname{grad} U = \Delta U = 0, \tag{13.134a}$$

i.e., to the *Laplace differential equation*. In Cartesian coordinates holds:

$$\Delta U = \frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} + \frac{\partial^2 U}{\partial z^2} = 0. \tag{13.134b}$$

Every function satisfying this differential equation and which is continuous and possesses continuous first and second order partial derivatives is called a *Laplace* or *harmonic function* (see also 14.1.2.2,**2.**, p. 732).

There are to distinguish three basic types of boundary value problems:

1. Boundary value problem (for an interior domain) or *Dirichlet problem*: A function $U(x, y, z)$ is determined, which is harmonic inside a given space or plane domain and takes the given values at the boundary of this domain.

2. Boundary value problem (for an interior domain) or *Neumann problem*: A function $U(x, y, z)$ is determined, which is harmonic inside a given domain and whose normal derivative $\dfrac{\partial U}{\partial n}$ takes the given values at the boundary of this domain.

3. Boundary value problem (for an interior domain): A function $U(x, y, z)$ is determined, which is harmonic inside a given domain and the expression $\alpha U + \beta \dfrac{\partial U}{\partial n}$ $\quad (\alpha, \beta \text{ const}, \quad \alpha^2 + \beta^2 \neq 0)$ takes the given values at the boundary of this domain.

13.5.2 Poisson Differential Equation

The problem to determine the potential U of a vector field $\vec{\mathbf{V}}_1 = \operatorname{grad} U$ with given divergence, leads to the equation according to (13.128) with $q(\vec{\mathbf{r}}) \neq 0$

$$\operatorname{div} \vec{\mathbf{V}}_{\mathbf{1}} = \operatorname{div} \operatorname{grad} U = \Delta U = q(\vec{\mathbf{r}}) \neq 0, \tag{13.135a}$$

i.e., to the *Poisson differential equation*. Since in Cartesian coordinates:

$$\Delta U = \frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} + \frac{\partial^2 U}{\partial z^2}, \tag{13.135b}$$

the Laplace differential equation (13.134b) is a special case of the Poisson differential equation (13.135b). The solution is the Newton potential (for point masses) or the Coulomb potential (for point charges)

$$U = -\frac{1}{4\pi} \iiint \frac{q(\vec{\mathbf{r}}^*)\, dv(\vec{\mathbf{r}}^*)}{|\vec{\mathbf{r}} - \vec{\mathbf{r}}^*|}. \tag{13.135c}$$

The integration is taken over the whole of space. $U(\vec{\mathbf{r}})$ tends to zero sufficiently rapidly for increasing $|\vec{\mathbf{r}}|$ values.

One can discuss the same three boundary value problems for the Poisson differential equation as for the solution of the Laplace differential equation in 13.5.1. The first and the third boundary value problems can be solved uniquely; for the second one there are to prescribe further special conditions (see [9.5]).

14 Function Theory

14.1 Functions of Complex Variables

14.1.1 Continuity, Differentiability

14.1.1.1 Definition of a Complex Function

Analogously to real functions, complex values can be assigned to complex values, i.e., to the value $z = x + \mathrm{i}\,y$ one can assign a complex number $w = u + \mathrm{i}\,v$, where $u = u(x, y)$ and $v = v(x, y)$ are real functions of two real variables. This relation is denoted by $w = f(z)$. The function $w = f(z)$ is a mapping from the complex z plane to the complex w plane.
The notions of limit, continuity, and derivative of a complex function $w = f(z)$ can be defined analogously to real functions of real variables.

14.1.1.2 Limit of a Complex Function

The *limit* of a function $f(z)$ is equal to the complex number w_0 if for z approaching z_0 the value of the function $f(z)$ approaches w_0:

$$w_0 = \lim_{z \to z_0} f(z). \tag{14.1a}$$

In other words: For any positive ε there is a (real) $\delta > 0$ such that for every z satisfying (14.1b), except maybe z_0 itself, the inequality (14.1c) holds:

$$|z_0 - z| < \delta, \tag{14.1b}$$

$$|w_0 - f(z)| < \varepsilon. \tag{14.1c}$$

The geometrical meaning is as follows: Any point z in the circle with center z_0 and radius δ, except maybe the center z_0 itself, is mapped into a point $w = f(z)$ inside a circle with center w_0 and radius ε in the w plane where f has its range, as shown in **Fig. 14.1**. The circles with radii δ and ε are also called the neighborhoods $U_\varepsilon(w_0)$ and $U_\delta(z_0)$.

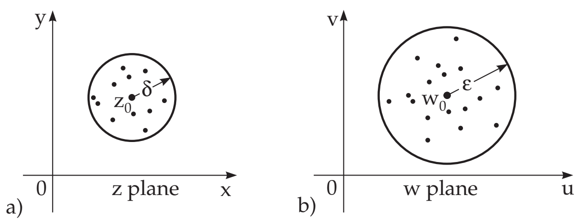


Figure 14.1

14.1.1.3 Continuous Complex Functions

A function $w = f(z)$ is continuous at z_0 if it has a limit there, and a substitution value, and these two are equal, i.e., if for an arbitrarily small given neighborhood $U_\varepsilon(w_0)$ of the point $w_0 = f(z_0)$ in the w plane there exists a neighborhood $U_\delta(z_0)$ of z_0 in the z plane such that $w = f(z) \in U_\varepsilon(w_0)$ for every $z \in U_\delta(z_0)$. As represented in **Fig. 14.1**, $U_\varepsilon(w_0)$ is, e.g., a circle with radius ε around the point w_0. The continuity of f is expressed by the usual notation

$$\lim_{z \to z_0} f(z) = f(z_0) \qquad \text{or} \qquad \lim_{\delta \to 0} f(z_0 + \delta) = f(z_0). \tag{14.2}$$

14.1.1.4 Differentiability of a Complex Function

A function $w = f(z)$ is differentiable at z if the difference quotient

$$\frac{\Delta w}{\Delta z} = \frac{f(z + \Delta z) - f(z)}{\Delta z} \tag{14.3}$$

I.N. Bronshtein et al., *Handbook of Mathematics*,
DOI 10.1007/978-3-662-46221-8_14

has a limit for $\Delta z \to 0$, independently of how Δz approaches zero. This limit is denoted by $f'(z)$ and it is called the derivative of $f(z)$.

■ The function $f(z) = \operatorname{Re} z = x$ is not differentiable at any point $z = z_0$, since approaching z_0 parallel to the x-axis the limit of the difference quotient is one, and approaching parallel to the y-axis this value is zero.

14.1.2 Analytic Functions

14.1.2.1 Definition of Analytic Functions

A function $f(z)$ is called *analytic*, *regular* or *holomorphic* on a domain G, if it is differentiable at every point of G. The boundary points of G, where $f'(z)$ does not exist, are singular points of $f(z)$.
The function $f(z) = u(x, y) + \mathrm{i}v(x, y)$ is differentiable in G if u and v have continuous partial derivatives in G with respect to x and y and they also satisfy the *Cauchy-Riemann differential equations*:

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}, \quad \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x}. \tag{14.4}$$

The real and imaginary parts of an analytic function satisfy the Laplace differential equation:

$$\Delta u(x, y) = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0, \tag{14.5a}$$

$$\Delta v(x, y) = \frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} = 0. \tag{14.5b}$$

The derivatives of the elementary functions of a complex variable can be calculated with the help of the same formulas as the derivative of the corresponding real functions.

■ **A:** $f(z) = z^3, \quad f'(z) = 3z^2$; ■ **B:** $f(z) = \sin z, \quad f'(z) = \cos z$.

14.1.2.2 Examples of Analytic Functions

1. Elementary Functions

The elementary algebraic and transcendental functions are analytic in the whole z plane except at some isolated singular points. If a function is analytic on a domain, i.e., it is differentiable, then it is differentiable arbitrarily many times.

■ **A:** The function $w = z^2$ with $u = x^2 - y^2, \ v = 2xy$ is everywhere analytic.

■ **B:** The function $w = u + \mathrm{i}v$, defined by the equations $u = 2x + y, \ v = x + 2y$, is not analytic at any point.

■ **C:** The function $f(z) = z^3$ with $f'(z) = 3z^2$ is analytic.

■ **D:** The function $f(z) = \sin z$ with $f'(z) = \cos z$ is analytic.

2. Determination of the Functions u and v

If both the functions u and v satisfy the Laplace differential equation, then they are *harmonic functions* (see 13.5.1, p. 729). If one of these harmonic functions is known, e.g., u, then the second one, as the conjugate harmonic function v, can be determined up to an additive constant with the Cauchy–Riemann differential equations:

$$v = \int \frac{\partial u}{\partial x}\,dy + \varphi(x) \quad \text{with} \quad \frac{d\varphi}{dx} = -\left(\frac{\partial u}{\partial y} + \frac{\partial}{\partial x}\int \frac{\partial u}{\partial x}\,dy\right). \tag{14.6}$$

Analogously u can be determined if v is known.

14.1.2.3 Properties of Analytic Functions

1. Absolute Value or Modulus of an Analytic Function

The absolute value (modulus) of an analytic function is:

$$|w| = |f(z)| = \sqrt{[u(x, y)]^2 + [v(x, y)]^2} = \varphi(x, y). \tag{14.7}$$

The surface $|w| = \varphi(x, y)$ is called its *relief*, i.e., $|w|$ is the third coordinate above every point $z = x + \mathrm{i}\,y$.

■ **A:** The absolute value of the function $\sin z = \sin x \cosh y + \mathrm{i} \cos x \sinh y$ is $|\sin z| = \sqrt{\sin^2 x + \sinh^2 y}$. The relief is shown in **Fig. 14.2a**.

■ **B:** The relief of the function $w = e^{1/z}$ is shown in **Fig. 14.2b**.

For the reliefs of several analytic functions see [14.10].

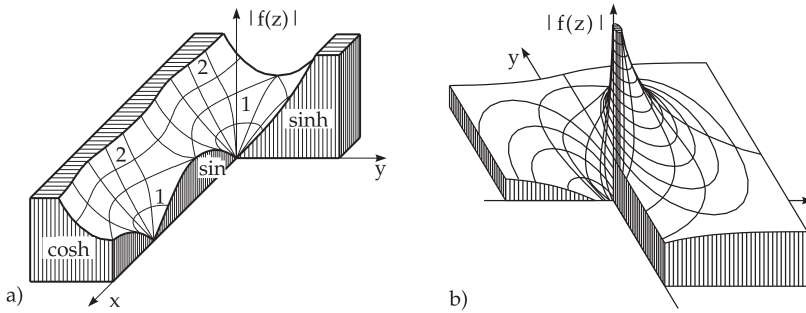


Figure 14.2

2. Roots

Since the absolute value of a function is positive or zero, the relief is always above the z plane, except the points where $|f(z)| = 0$ holds, so $f(z) = 0$. The z values, where $f(z) = 0$ holds, are called the *roots of the function* $f(z)$.

3. Boundedness

A function $f(z)$ is *bounded* on a certain domain G, if there exists a positive number N such that $|f(z)| < N$ for all z in G. In the opposite case, if no such number N exists, then the function is called unbounded in G.

4. Theorem about the Maximum Value

If $w = f(z)$ is an analytic function on a closed domain, then the maximum of its absolute value is attained on the boundary of the domain.

5. Theorem about the Constant (*Theorem of Liouville*)

If $w = f(z)$ is analytic in the whole z plane and also bounded, then this function is a constant: $f(z) =$ const.

14.1.2.4 Singular Points

If a function $w = f(z)$ is analytic in a neighborhood of $z = a$, i.e., in a small circle with center a, except a itself, then f has a singularity at a. There exist three types of singularities:

1. $f(z)$ is bounded in the neighborhood. Then there exists $w = \lim\limits_{z \to a} f(z)$, and setting $f(a) = w$ the function becomes analytic also at a. In this case, f has a *removable singularity* at a.

2. If $\lim\limits_{z \to a} |f(z)| = \infty$, then f has a pole at a. About poles of different orders see 14.3.5.1, p. 753.

3. If f has neither a removable singularity nor a pole at a, then f has an *essential singularity* at a. In this case, for any complex w there exists a sequence $z_n \to a$ such that $f(z_n) \to w$.

■ **A:** The function $w = \dfrac{1}{z - a}$ has a pole at a.

■ **B:** The function $w = e^{1/z}$ has an essential singularity at 0 **(Fig. 14.2b)**.

14.1.3 Conformal Mapping

14.1.3.1 Notion and Properties of Conformal Mappings

1. Definition

A mapping from the z plane to the w plane is called a conformal mapping if it is analytic and injective. In this case,

$$w = f(z) = u + \mathrm{i}v, \quad f'(z) \neq 0. \tag{14.8}$$

The conformal mapping has the following properties:

The transformation $dw = f'(z)\,dz$ of the line element $dz = \begin{pmatrix} dx \\ dy \end{pmatrix}$ is the composition of a dilatation by $\sigma = |f'(z)|$ and of a rotation by $\alpha = \arg f'(z)$. This means that infinitesimal circles are transformed into almost circles, triangles into (almost) similar triangles **(Fig. 14.3)**. The curves keep their angles of intersection, so an orthogonal family of curves is transformed into an orthogonal family **(Fig. 14.4)**.

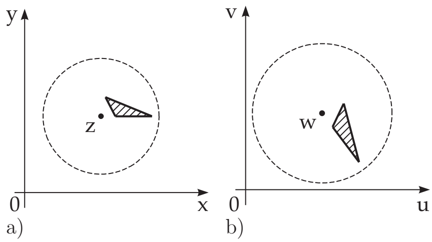


Figure 14.3

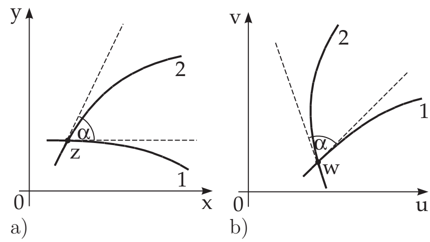


Figure 14.4

Remark: Conformal mappings can be found in physics, electrotechnics, hydro- and aerodynamics and in other areas of mathematics.

2. The Cauchy–Riemann Equations

The mapping between dz and dw is given by the affine differential transformation

$$du = \frac{\partial u}{\partial x}\,dx + \frac{\partial u}{\partial y}\,dy, \qquad dv = \frac{\partial v}{\partial x}\,dx + \frac{\partial v}{\partial y}\,dy \tag{14.9a}$$

and in matrix form

$$dw = \mathbf{A}\,dz \qquad \text{with} \qquad \mathbf{A} = \begin{pmatrix} u_x & u_y \\ v_x & v_y \end{pmatrix}. \tag{14.9b}$$

According to the Cauchy–Riemann differential equations, $\mathbf{A}$ has the form of a rotation-stretching matrix (see 3.5.2.2, **2.**, p. 192) with σ as the stretching factor:

$$\mathbf{A} = \begin{pmatrix} u_x & -v_x \\ v_x & u_x \end{pmatrix} = \sigma \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix}, \tag{14.10a}$$

$$u_x = v_y = \sigma\cos\alpha\,, \tag{14.10b}$$

$$\sigma = |f'(z)| = \sqrt{u_x^2 + u_y^2} = \sqrt{v_x^2 + v_y^2}\,, \tag{14.10c}$$

$$-u_y = v_x = \sigma\sin\alpha\,, \tag{14.10d}$$

$$\alpha = \arg f'(z) = \arg\,(u_x + \mathrm{i}v_x)\,. \tag{14.10e}$$

3. Orthogonal Systems

The coordinate lines $x =$ const and $y =$ const of the z plane are transformed by a conformal mapping into two orthogonal families of curves. In general, a bunch of orthogonal curvilinear coordinate systems

can be generated by analytic functions; and conversely, for every conformal mapping there exist an orthogonal net of curves which is transformed into an orthogonal coordinate system.

■ **A:** In the case of $u = 2x + y,\ v = x + 2y$ **(Fig. 14.5)**, orthogonality does not hold.

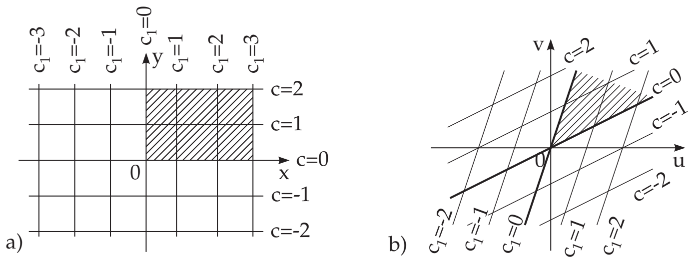


Figure 14.5

■ **B:** In the case $w = z^2$ the orthogonality is retained, except at the point $z = 0$ because here $w' = 0$. The coordinate lines are transformed into two confocal families of parabolas **(Fig. 14.6)**, the first quadrant of the z plane into the upper half of the w plane.

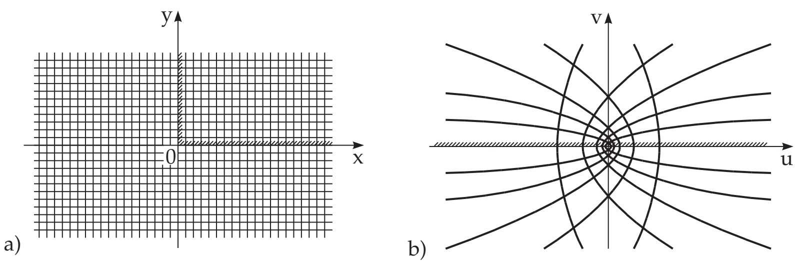


Figure 14.6

14.1.3.2 Simplest Conformal Mappings

In this paragraph, some transformations are discussed with their most important properties, and there are given the graphs of *isometric nets* in the z plane, i.e., the nets which are transformed into an orthogonal Cartesian net in the w plane. The boundaries of the z regions mapped into the upper half of the w plane are denoted by shading. Black regions are mapped by the conformal mapping onto the square in the w plane with vertices $(0,0)$, $(0,1)$, $(1,0)$, and $(1,1)$ **(Fig. 14.7)**.

1. Linear Function

For the conformal mapping given in the form of a linear function

$$w = az + b \quad (a, b \text{ complex, const}; a \neq 0), \tag{14.11a}$$

the transformation can be done in three steps:

a) Rotation of the z plane by the angle $\alpha = \arg a$ according to : $\quad w_1 = e^{i\alpha} z.$

b) Stretching of the w_1 plane by the factor $|a|$: $\quad w_2 = |a| w_1.$ (14.11b)

c) Parallel translation of the w_2 plane by b : $\quad w = w_2 + b.$

Altogether, every figure of the z plane is transformed into a similar one of the w plane. The points $z_1 = \infty$ and $z_2 = \dfrac{b}{1-a}$ for $a \neq 1$ are transformed into themselves, and they are called *fixed points*.

Fig. 14.8 shows the orthogonal net which is transformed into the orthogonal Cartesian net.

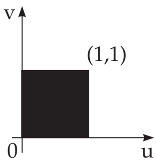


Figure 14.7

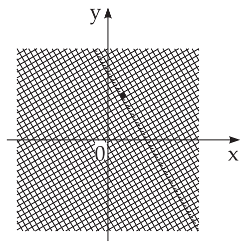


Figure 14.8

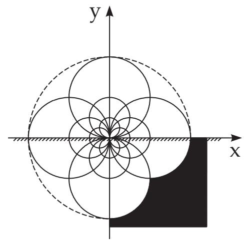


Figure 14.9

2. Inversion

The conformal mapping

$$w = \frac{1}{z} \tag{14.12}$$

represents an *inversion* with respect to the *unit circle* and a reflection in the real axis, namely, a point $z = re^{i\varphi}$ of the z plane with the absolute value r and with the argument φ is transformed into a point $w = \frac{1}{r}e^{-i\varphi}$ of the w plane with the absolute value $1/r$ and with the argument $-\varphi$ (see **Fig. 14.10**). Circles are transformed into circles, where lines are considered as limiting cases of circles (radius $\to \infty$). Points of the interior of the unit circle become exterior points and conversely (see **Fig. 14.11**). The point $z = 0$ is transformed into $w = \infty$. The points $z = -1$ and $z = 1$ are *fixed points* of this conformal mapping. The orthogonal net of the transformation (14.12) is shown in **Fig. 14.9**.

Remark: In general a geometric transformation is called *inversion with respect to a circle with radius* r, when a point P_2 with radius r_2 inside the circle is transformed into a point P_1 on the elongation of the same radius vector $\overrightarrow{OP_2}$ outside of the circle with radius $\overline{OP}_1 = r_1 = r^2/r_2$. Points of the interior of the circle become exterior points and conversely.

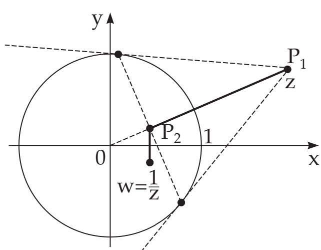


Figure 14.10

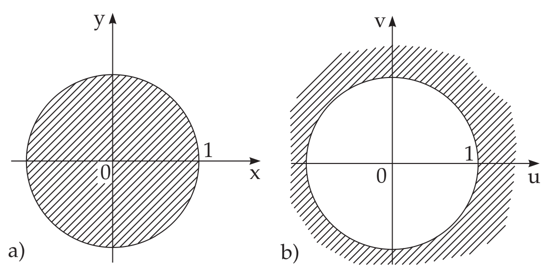


Figure 14.11

3. Linear Fractional Function

For the conformal mapping given in the form of a linear fractional function

$$w = \frac{az+b}{cz+d} \quad (a, b, c, d \text{ complex, const}; bc - ad \neq 0; c \neq 0) \tag{14.13a}$$

the transformation can be performed in three steps:

$$\begin{aligned} &\textbf{a)}\ \text{Linear function:} && w_1 = cz + d. \\ &\textbf{b)}\ \text{Inversion:} && w_2 = \frac{1}{w_1}. \\ &\textbf{c)}\ \text{Linear function:} && w = \frac{a}{c} + \frac{bc - ad}{c} w_2. \end{aligned} \tag{14.13b}$$

Circles are transformed again into circles (*circular transformation*), where straight lines are considered as limiting cases of circles with radius $r \to \infty$. Fixed points of this conformal mapping are the both points satisfying the quadratic equation

$$z = \frac{az + b}{cz + d}. \tag{14.14}$$

If the points z_1 and z_2 are inverses of each other with respect to a circle K_1 of the z plane, then their images w_1 and w_2 in the w plane are also inversions of each other with respect to the image circle K_2 of K_1.
The orthogonal net which has the orthogonal Cartesian net as its image is represented in **Fig. 14.12**.

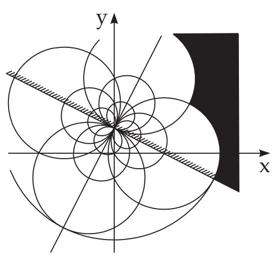


Figure 14.12

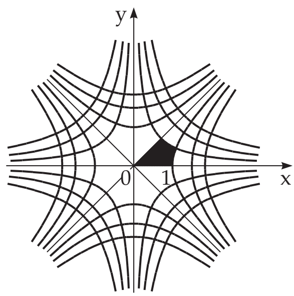


Figure 14.13

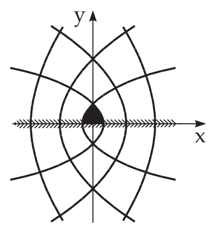


Figure 14.14

4. Quadratic Function

The conformal mapping described by a quadratic function

$$w = z^2 \tag{14.15a}$$

has the form in polar coordinates and as a function of x and y:

$$w = \rho^2 e^{i2\varphi}, \tag{14.15b}$$ $$w = u + i\,v = x^2 - y^2 + 2i\,xy. \tag{14.15c}$$

It is obvious from the polar coordinate representation that the upper half of the z plane is mapped onto the whole w plane, i.e., the whole image of the z plane will cover twice the whole w plane.
The representation in Cartesian coordinates shows that the coordinate lines of the w plane, $u =$ const and $v =$ const, come from the orthogonal families of hyperbolas $x^2 - y^2 = u$ and $2xy = v$ of the z plane **(Fig. 14.13)**.
Fixed points of this mapping are $z = 0$ and $z = 1$. This mapping is not conformal at $z = 0$.

5. Square Root

The conformal mapping given in the form as a square root of z,

$$w = \sqrt{z}\,, \tag{14.16}$$

transforms the whole z plane whether onto the upper half of the w plane or onto the lower half of it, i.e., the function is double-valued, i.e., to every value of z ($z \neq 0$) belong two values of w (see 1.5.3.6, p. 38). The coordinate lines of the w plane come from two orthogonal families of confocal parabolas with the focus at the origin of the z plane and with the positive or with the negative real half-axis as

their axis **(Fig. 14.14)**.
Fixed points of the mapping are $z = 0$ and $z = 1$. The mapping is not conformal at $z = 0$.

6. Sum of Linear and Fractional Linear Functions

The conformal mapping given by the function

$$w = \frac{k}{2}\left(z + \frac{1}{z}\right) \quad (k \text{ a real constant}, \quad k > 0) \tag{14.17a}$$

can be transformed by the polar coordinate representation of $z = \rho\, e^{\mathrm{i}\varphi}$ and by separating the real and imaginary parts according to (14.8):

$$u = \frac{k}{2}\left(\rho + \frac{1}{\rho}\right)\cos\varphi, \quad v = \frac{k}{2}\left(\rho - \frac{1}{\rho}\right)\sin\varphi. \tag{14.17b}$$

Circles with $\rho = \rho_0 = \text{const}$ of the z plane **(Fig. 14.15a)** are transformed into confocal ellipses

$$\frac{u^2}{a^2} + \frac{v^2}{b^2} = 1 \quad \text{with} \quad a = \frac{k}{2}\left(\rho_0 + \frac{1}{\rho_0}\right), \quad b = \frac{k}{2}\left|\rho_0 - \frac{1}{\rho_0}\right| \tag{14.17c}$$

in the w plane **(Fig. 14.15b)**. The foci are the points $\pm k$ of the real axis. For the unit circle with $\rho = \rho_0 = 1$ one gets the degenerated ellipse of the w plane, the twice overrunning segment $(-k, +k)$ of the real axis. Both the interior and the exterior of the unit circle are transformed onto the entire w plane with the cut $(-k, +k)$, so its inverse function is double-valued:

$$z = \frac{w + \sqrt{w^2 - k^2}}{k}. \tag{14.17d}$$

The lines $\varphi = \varphi_0$ of the z plane **(Fig. 14.15c)** become confocal hyperbolas

$$\frac{u^2}{\alpha^2} - \frac{v^2}{\beta^2} = 1 \quad \text{with} \quad \alpha = k\cos\varphi_0, \quad \beta = k\sin\varphi_0 \tag{14.17e}$$

with foci $\pm k$ **(Fig. 14.15d)**. The hyperbolas corresponding to the coordinate half-axis of the z plane $\left(\varphi = 0, \frac{\pi}{2}, \pi, \frac{3}{2}\pi\right)$ are degenerate in the axis $u = 0$ (v axis) and in the intervals $(-\infty, -k)$ and (k, ∞) of the real axis running there and back.

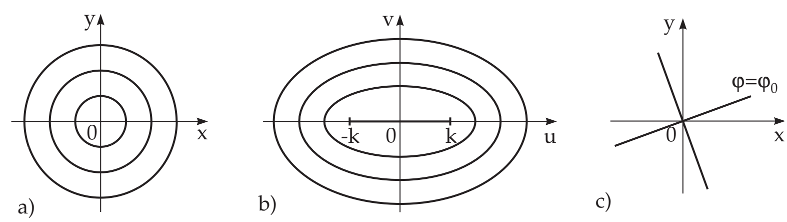


Figure 14.15

7. Logarithm

The conformal mapping given in the form of the logarithm function

$$w = \operatorname{Ln} z \tag{14.18a}$$

has the form for z given in polar coordinates:

$$u = \ln\rho, \quad v = \varphi + 2k\pi \qquad (k = 0, \pm 1, \pm 2, \ldots). \tag{14.18b}$$

One can see from this representation that the coordinate lines $u = \text{const}$ and $v = \text{const}$ come from concentric circles around the origin of the z plane and from rays starting at the origin of the z plane

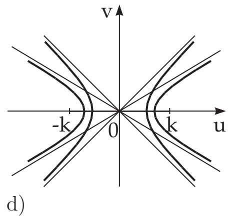


Figure 14.15

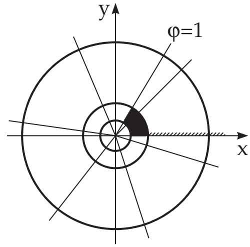


Figure 14.16

(Fig. 14.16). The isometric net is a polar net.
The logarithm function $\operatorname{Ln} z$ is infinitely many valued (see (14.74c), p. 758).
Restricting the investigation to the principal value $\ln z$ of $\operatorname{Ln} z$ $(-\pi < v \leq +\pi)$, then the whole z plane is transformed into a stripe of the w plane bounded by the lines $v = \pm\pi$, where $v = \pi$ belongs to the stripe.

8. Exponential Function

The conformal mapping given in the form of an exponential function (see also 14.5.2, **1.**, p. 758)

$$w = e^z \tag{14.19a}$$

has the form in polar coordinates:

$$w = \rho e^{i\psi}. \tag{14.19b}$$

From $z = x + i\,y$ follows:

$$\rho = e^x \quad \text{and} \quad \psi = y. \tag{14.19c}$$

If y changes from $-\pi$ to $+\pi$, and x changes from $-\infty$ to $+\infty$, then ρ takes all values from 0 to ∞ and ψ from $-\pi$ to π. A 2π wide stripe of the z plane, parallel to the x-axis, will be transformed into the entire w plane **(Fig. 14.17)**.

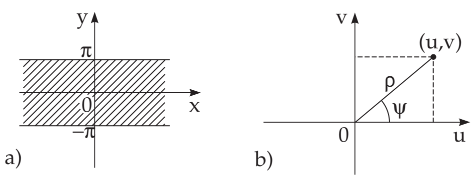


Figure 14.17

9. The Schwarz-Christoffel Formula

By the Schwarz-Christoffel formula

$$z = C_1 \int_0^w \frac{dt}{(t-w_1)^{\alpha_1}(t-w_2)^{\alpha_2}(t-w_n)^{\alpha_n}} + C_2 \tag{14.20a}$$

the interior of a polygon of the z plane can be mapped onto the upper half of the w plane. The polygon has n exterior angles $\alpha_1\pi$, $\alpha_2\pi, \ldots,$ $\alpha_n\pi$ **(Fig. 14.18a,b)**. The points of the real axis in the w plane assigned to the vertices of the polygon are denoted by w_i $(i = 1, \ldots, n)$, and the integration variable is denoted by t. The oriented boundary of the polygon is transformed into the oriented real axis of the w

plane by this mapping. For large values of $|t|$, the integrand behaves as $1/t^2$ and is regular at infinity. Since the sum of all the exterior angles of an n-gon is equal to 2π, consequently

$$\sum_{\nu=1}^{n} \alpha_\nu = 2. \tag{14.20b}$$

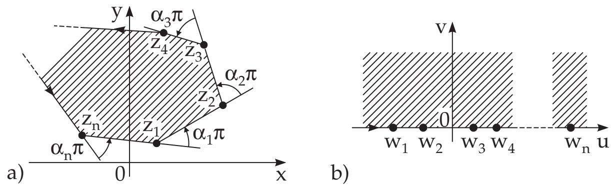


Figure 14.18

The complex constants C_1 and C_2 yield a rotation, a stretching and a translation; they do not depend on the form of the polygon, only on the size and the position of the polygon in the z plane.

Three arbitrary points w_1, w_2, w_3 of the w plane can be assigned to three points z_1, z_2, z_3 of the polygon in the z plane. Assigning a point at infinity in the w plane, i.e., $w_1 = \pm\infty$ to a vertex of the polygon in the z plane, e.g., to $z = z_1$, then the factor $(t - w_1)^{\alpha_1}$ is omitted. If the polygon is degenerated, e.g., a vertex is at infinity, then the corresponding exterior angle is π, so $\alpha_\infty = 1$, i.e., the polygon is a half-stripe.

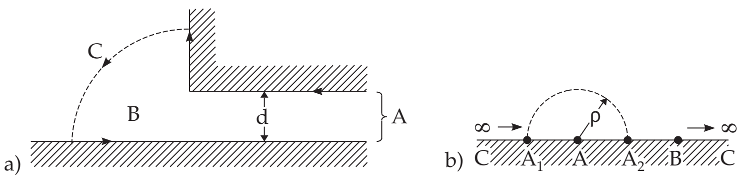


Figure 14.19

■ **A:** Mapping of a certain region of the z plane (shaded region in **Fig. 14.19a**): Considering $\sum \alpha_\nu = 2$ one chooses three points as the table shows **(Fig. 14.19a,b)**. The formula of the mapping is:

	z_ν	α_ν	w_ν
A	∞	1	-1
B	0	$-1/2$	0
C	∞	$3/2$	∞

$$z = C_1 \int_0^w \frac{dt}{(t+1)t^{-1/2}} = 2C_1\left(\sqrt{w} - \arctan\sqrt{w}\right) = \mathrm{i}\frac{2d}{\pi}\left(\sqrt{w} - \arctan\sqrt{w}\right).$$

To determinate C_1 one substitutes $t = \rho e^{\mathrm{i}\varphi} - 1$, giving $\mathrm{i}d = C_1 \lim\limits_{\rho\to 0} \displaystyle\int_\pi^0 \frac{\left(-1+\rho e^{\mathrm{i}\varphi}\right)^{1/2} \mathrm{i}\rho e^{\mathrm{i}\varphi} d\varphi}{\rho e^{\mathrm{i}\varphi}} = C_1\pi$,

i.e., $C_1 = \mathrm{i}\dfrac{d}{\pi}$.

For the constant C_2 follows $C_2 = 0$, considering that the mapping assigns "$z = 0 \to w = 0$".

■ **B:** Mapping of a rectangle. Let the vertices of the rectangle be $z_{1,4} = \pm K,\ z_{2,3} = \pm K + \mathrm{i}K'$. The points z_1 and z_2 should be transformed into the points $w_1 = 1$ and $w_2 = 1/k$ $(0 < k < 1)$ of the real axis, z_4 and z_3 are reflections of z_1 and z_2 with respect to the imaginary axis. According to the Schwarz reflection principle (see 14.1.3.3, p. 741) they must correspond to the points $w_4 = -1$ and $w_3 = -1/k$ **(Fig. 14.20a,b)**. So, the mapping formula for a rectangle ($\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 1/2$) of the position sketched above is: $z = C_1 \displaystyle\int_0^w \frac{dt}{\sqrt{(t-w_1)(t-w_2)(t-w_3)(t-w_4)}} = C_1 \int_0^w \frac{dt}{\sqrt{(t^2-1)\left(t^2 - \frac{1}{k^2}\right)}}$. The

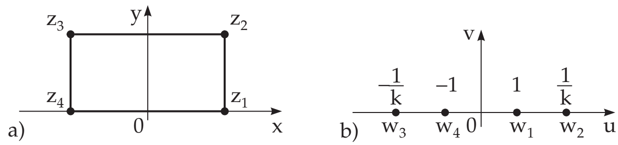


Figure 14.20

point $z = 0$ has the image $w = 0$ and the image of $z = \mathrm{i}K$ is $w = \infty$. With $C_1 = 1/k$ follows

$$z = \int_0^w \frac{dt}{\sqrt{(1-t^2)(1-k^2t^2)}} = \int_0^\varphi \frac{d\vartheta}{\sqrt{1-k^2\sin^2\vartheta}} = F(\varphi, k) \quad (\text{substituting } t = \sin\vartheta,\ w = \sin\varphi).$$

$F(\varphi, k)$ is the elliptic integral of the first kind (see 8.1.4.3, p. 490).
For the constant C_2 one gets $C_2 = 0$, considering that the mapping assigns " $z = 0 \to w = 0$ ".

14.1.3.3 Schwarz Reflection Principle

1. Statement

Suppose $f(z)$ is an analytic complex function in a domain G, and the boundary of G contains a line segment g_1. If the function is continuous on g_1 and it maps the line g_1 into a line g_1', then the points lying symmetric with respect to g_1 are transformed into points lying symmetric with respect to g_1' **(Fig. 14.21)**.

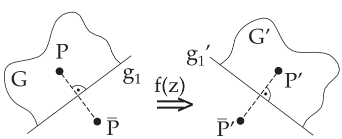


Figure 14.21

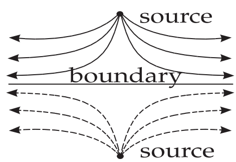


Figure 14.22

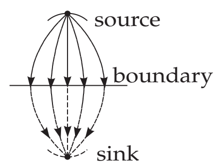


Figure 14.23

2. Application

The application of this principle makes it easier to perform calculations and the representations of plane regions with straight line boundaries: If the line boundary is a stream line (isolating boundary in **Fig. 14.22**), then the sources are reflected as sources, the sinks as sinks and curls as curls with the opposite sense of rotation. If the line boundary is a potential line (heavy conducting boundary in **Fig. 14.23**), then the sources are reflected as sinks, the sinks as sources and curls as curls with the same sense of rotation.

14.1.3.4 Complex Potential

1. Notion of the Complex Potential

A field $\vec{\mathbf{V}} = \vec{\mathbf{V}}(x, y)$ is considered in the x, y plane for the zero-divergence and the irrotational case with continuous and differentiable components $V_x(x, y)$ and $V_y(x, y)$ of the vector $\vec{\mathbf{V}}$.

a) Zero-divergence field with $\operatorname{div} \vec{\mathbf{V}} = 0$, i.e., $\dfrac{\partial V_x}{\partial x} + \dfrac{\partial V_y}{\partial y} = 0$: That is the integrability condition for the differential equation expressed with the *field* or *stream function* $\Psi(x, y)$

$$d\Psi = -V_y\, dx + V_x\, dy = 0, \quad (14.21a) \quad \text{and then} \quad V_x = \frac{\partial \Psi}{\partial y}, \qquad V_y = -\frac{\partial \Psi}{\partial x}. \qquad (14.21b)$$

For two points P_1, P_2 of the field $\vec{\mathbf{V}}$ the difference $\Psi(P_2) - \Psi(P_1)$ is a measure of the flux through a curve connecting the points P_1 and P_2, in the case when the whole curve is in the field.

b) Irrotational field with rot $\vec{\mathbf{V}} = \vec{\mathbf{0}}$, i.e., $\dfrac{\partial V_y}{\partial x} - \dfrac{\partial V_x}{\partial y} = 0$: That is the integrability condition for the differential equation with the potential function $\Phi(x, y)$

$$d\Phi = V_x\,dx + V_y\,dy = 0, \quad (14.22a) \qquad \text{and then} \quad V_x = \frac{\partial \Phi}{\partial x}, \qquad V_y = \frac{\partial \Phi}{\partial y}. \tag{14.22b}$$

If the field is free of sources and rotations, then the functions Φ and Ψ satisfy the Cauchy–Riemann differential equations (see 14.1.2.1, p. 732) and both satisfy the Laplace differential equation ($\Delta\Phi = 0$, $\Delta\Psi = 0$). Combining the functions Φ and Ψ into the analytic function

$$W = f(z) = \Phi(x, y) + \mathrm{i}\,\Psi(x, y), \tag{14.23}$$

then this function is called the complex potential of the field $\vec{\mathbf{V}}$.
Then $-\Phi(x, y)$ is the potential of the vector field $\vec{\mathbf{V}}$ in the sense of the usual notation in physics and electrotechnics (see 13.3.1.6, **2.**, p. 721). The level lines of Ψ and Φ form an orthogonal net. For the derivation of the complex potential and the field vector $\vec{\mathbf{V}}$ the following equations hold:

$$\frac{dW}{dz} = \frac{\partial \Phi}{\partial x} - \mathrm{i}\frac{\partial \Phi}{\partial y} = V_x - \mathrm{i}V_y, \qquad \overline{\frac{dW}{dz}} = \overline{f'(z)} = V_x + \mathrm{i}V_y. \tag{14.24}$$

2. Complex Potential of a Homogeneous Field

The function

$$W = a\,z, \tag{14.25}$$

with real a is the complex potential of a field whose potential lines are parallel to the y-axis and whose direction lines are parallel to the x-axis **(Fig. 14.24)**. A complex a results in a rotation of the field **(Fig. 14.25)**.

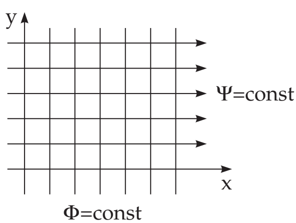


Figure 14.24

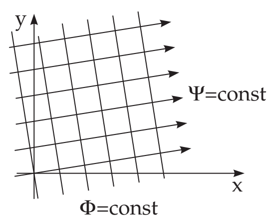


Figure 14.25

3. Complex Potential of Source and Sink

The complex potential of a field with a strength of source $s > 0$ at the point $z = z_0$ has the equation:

$$W = \frac{s}{2\pi}\ln(z - z_0). \tag{14.26}$$

A sink with the same intensity has the equation:

$$W = -\frac{s}{2\pi}\ln(z - z_0). \tag{14.27}$$

The direction lines run away radially from $z = z_0$, while the potential lines are concentric circles around the point z_0 **(Fig. 14.26)**.

4. Complex Potential of a Source–Sink System

One gets the complex potential for a source at the point z_1 and for a sink at the point z_2, both having the same intensity, by superposition:

$$W = \frac{s}{2\pi} \ln \frac{z - z_1}{z - z_2}. \tag{14.28}$$

The potential lines Φ = const form *Apollonius circles* with respect to z_1 and z_2; the direction lines Ψ = const are circles through z_1 and z_2 **(Fig. 14.27)**.

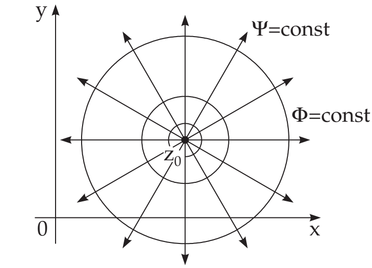


Figure 14.26

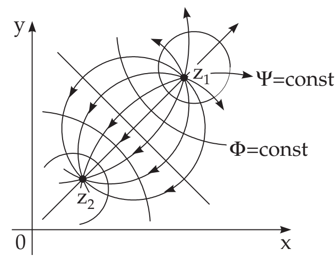


Figure 14.27

5. Complex Potential of a Dipole

The complex potential of a dipole with dipole moment $M > 0$ at the point z_0, whose axis encloses an angle α with the real axis **(Fig. 14.28)**, has the equation:

$$W = \frac{M e^{i\alpha}}{2\pi(z - z_0)}. \tag{14.29}$$

6. Complex Potential of a Curl

If the intensity of the curl is $|\Gamma|$ with real Γ and its center is at the point z_0, then its equation is:

$$W = \frac{\Gamma}{2\pi i} \ln(z - z_0). \tag{14.30}$$

In comparison with **Fig. 14.26**, the roles of the direction lines and the potential are interchanged. For complex Γ (14.30) gives the potential of a source of curl, whose direction and potential lines form two families of spirals orthogonal to each other **(Fig. 14.29)**.

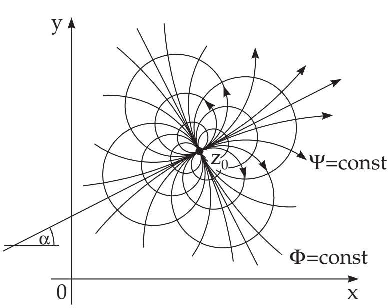


Figure 14.28

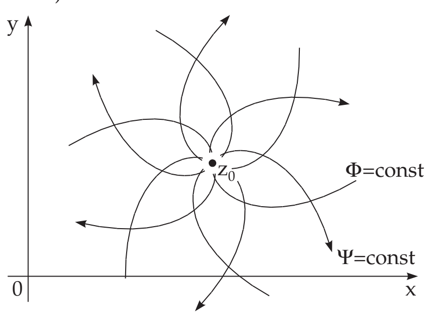


Figure 14.29

14.1.3.5 Superposition Principle

1. Superposition of Complex Potentials

A system composed of several sources, sinks, and curls is an additive superposition of single fields, i.e., one gets its function by adding their complex potential and stream functions. Mathematically this is possible because of the linearity of the Laplace differential equations $\Delta\Phi = 0$ and $\Delta\Psi = 0$.

2. Composition of Vector Fields

1. Integration Beside addition from complex potentials new fields can be constructed also by integration with the application of weight functions.

■ A curl-covering is given on a line segment l with density function $\varrho(s)$. The derivative of the complex potential is given by an integral of Cauchy-type (see 14.2.3, p. 748)

$$\frac{dW}{dz} = \frac{1}{2\pi\mathrm{i}}\int_{(l)} \frac{\varrho(s)\,ds}{z-\zeta(s)} = \frac{1}{2\pi\mathrm{i}}\int_{(l)} \frac{\varrho^*(\zeta)}{z-\zeta}\,d\zeta\,, \tag{14.31}$$

where $\zeta(s)$ is the complex parameter representation of the curve l with arc length parameter s.

2. Maxwell Diagonal Method If one wants to compose the superposition of two fields with the potentials Φ_1 and Φ_2, then one can draw the potential line figures $[[\Phi_1]]$ and $[[\Phi_2]]$ so that the value of the potential changes by the same amount h between two neighboring lines in both systems. Then the lines are directed so that the higher Φ values are on the left-hand side. The lines lying in the diagonal direction to the net elements formed by $[[\Phi_1]]$ and $[[\Phi_2]]$ give the potential lines of the field $[[\Phi]]$, whose potential is $\Phi = \Phi_1 + \Phi_2$ or $\Phi = \Phi_1 - \Phi_2$. One gets the figure of $[[\Phi_1 + \Phi_2]]$ if the oriented sides of the net elements are added as vectors **(Fig. 14.30a)**, and one gets the figure of $[[\Phi_1 - \Phi_2]]$ by subtracting them **(Fig. 14.30b)**. The value of the composed potential changes by h at transition from one potential line to the next one.

■ Vector lines and potential lines of a source and a sink with an intensity quotient $|e_1|/|e_2| = 3/2$ (see **Fig. 14.31a,b**).

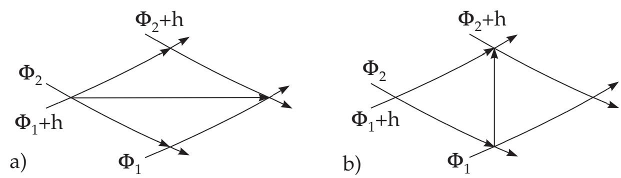


Figure 14.30

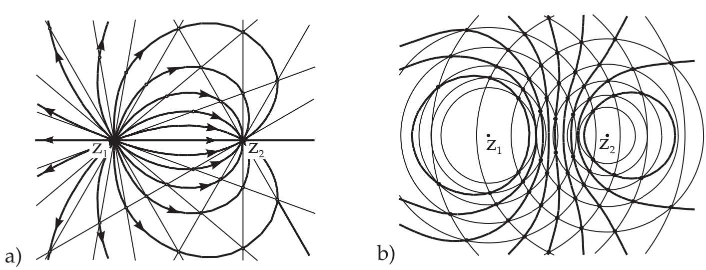


Figure 14.31

14.1.3.6 Arbitrary Mappings of the Complex Plane

A function

$$w = f(z = x + \mathrm{i}\, y) = u(x, y) + \mathrm{i}\, v(x, y) \tag{14.32a}$$

is defined if the two functions $u = u(x, y)$ and $v = v(x, y)$ with real variables are defined and known. The function $f(z)$ must not be analytic, as it was required in conformal mappings. The function w maps the z plane into the w plane, i.e., it assigns to every point z_ν a corresponding point w_ν.

a) Transformation of the Coordinate Lines

$$y = c \longrightarrow u = u(x, c), \quad v = v(x, c), \quad x \text{ is a parameter;}$$
$$x = c_1 \longrightarrow u = u(c_1, y), \quad v = v(c_1, y), \quad y \text{ is a parameter.} \tag{14.32b}$$

b) Transformation of Geometric Figures Geometric figures as curves or regions of the z plane are usually transformed into curves or regions of the w plane:

$$x = x(t),\ \ y = y(t) \quad \to \quad u = u(x(t), y(t)),\ \ v = v(x(t), y(t)), \quad t \text{ is a parameter.} \tag{14.32c}$$

■ For $u = 2x + y$, $v = x + 2y$, the lines $y = c$ are transformed into $u = 2x + c$, $v = x + 2c$, hence into the lines $v = \dfrac{u}{2} + \dfrac{3}{2}c$. The lines $x = c_1$ are transformed into the lines $v = 2u - 3c_1$ **(Fig. 14.5)**. The shaded region in **Fig. 14.5a** is transformed into the shaded region in **Fig. 14.5b**.

c) Riemann Surface Getting the same value w for several different z of the mapping $w = f(z)$, then the image of the function consists of several planes "lying on each other". Cutting these planes and connecting them along a curve gives a many-sheeted surface, the so-called *many-sheeted Riemann surface* (see [14.13]). This correspondence can be considered also in an inverse relation, in the case of multiple-valued functions as, e.g., the functions $\sqrt[n]{z}$, $\operatorname{Ln} z$, $\operatorname{Arcsin} z$, $\operatorname{Arctan} z$.

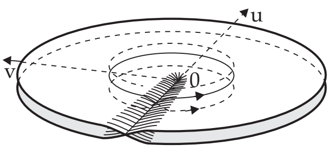


Figure 14.32

■ $w = z^2$: While $z = re^{\mathrm{i}\varphi}$ overruns the entire z plane, i.e., $0 \le \varphi < 2\pi$, the values of $w = \varrho e^{\mathrm{i}\psi} = r^2 e^{\mathrm{i}2\varphi}$, cover twice the w plane. One can imagine to place two w planes one on each other, and cut and connect them along the negative real axis according to **Fig. 14.32**. This surface is called the Riemann surface of the function $w = z^2$. The zero point is called a *branch point*. The image of the function e^z (see 14.69) is a Riemann surface of infinitely many sheets.

(In many cases the planes are cut along the positive real axis. It depends on whether the principal value of the complex number is defined for the interval $(-\pi, +\pi]$ or for the interval $[0, 2\pi)$.)

14.2 Integration in the Complex Plane

14.2.1 Definite and Indefinite Integral

14.2.1.1 Definition of the Integral in the Complex Plane

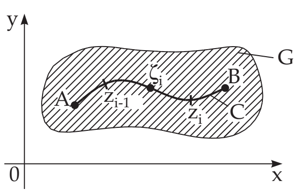


Figure 14.33

1. Definite Complex Integral

Suppose $f(z)$ is continuous in a domain G, and the curve C is rectifiable, it connects the points A and B, and the whole curve is in this domain. The curve C is decomposed between the points A and B by arbitrary division points z_i into n subarcs **(Fig.14.33)**.

A point ζ_i is chosen on every arc segment and forms the sum

$$\sum_{i=1}^{n} f(\zeta_i)\, \Delta z_i \quad \text{with} \quad \Delta z_i = z_i - z_{i-1}. \tag{14.33a}$$

If the limit

$$\lim_{n\to\infty}\sum_{i=1}^{n} f(\zeta_i)\,\Delta z_i \tag{14.33b}$$

exists for $\Delta z_i \to 0$ and $n \to \infty$ independently of the choice of the points z_i and ζ_i, then this limit is called the *definite complex integral*

$$I = \int\limits_{\widehat{AB}} f(z)\,dz = (C)\int\limits_{A}^{B} f(z)\,dz \tag{14.33c}$$

along the curve C between the points A and B. The value of the integral usually depends on the path of the integral.

2. Indefinite Complex Integral

If the definite integral is independent of the path of the integral (see 14.2.2, p. 747), then holds:

$$F(z) = \int f(z)\,dz + C \quad \text{with} \quad F'(z) = f(z). \tag{14.34}$$

Here C is the integration constant which is complex, in general. The function $F(z)$ is called an *indefinite complex integral.*

The indefinite integrals of the elementary functions of a complex variable can be calculated with the same formulas as the integrals of the corresponding elementary function of a real variable.

■ **A:** $\int \sin z\,dz = -\cos z + C.$ ■ **B:** $\int e^z\,dz = e^z + C.$

3. Relation between Definite and Indefinite Complex Integrals

If the function $f(z)$ is analytic (see 14.1.2.1, p. 732) and has an indefinite integral, then the relation between its definite and indefinite integral is

$$\int\limits_{\widehat{AB}} f(z)\,dz = \int\limits_{A}^{B} f(z)\,dz = F(z_B) - F(z_A). \tag{14.35}$$

14.2.1.2 Properties and Evaluation of Complex Integrals

1. Comparison with the curvilinear integral of the second type

The definite complex integral has the same properties as the curvilinear integral of the second type (see 8.3.2, p. 517):

a) Reversing the direction of the path of integration the integral changes its sign.

b) Decomposing the path of integration into several parts the value of the total integral is the sum of the integrals on the parts.

2. Estimation of the Value of the Integral

If the absolute value of the function $f(z)$ does not exceed a positive number M at the points z of the path of integration $\widehat{AB}$ which has the length s, then:

$$\left|\int\limits_{\widehat{AB}} f(z)\,dz\right| \le Ms \quad \text{if} \quad |f(z)| \le M. \tag{14.36}$$

3. Evaluation of the Complex Integral in Parametric Representation

If the path of integration $\widehat{AB}$ (or the curve C) is given in the form

$$x = x(t), \qquad y = y(t) \tag{14.37}$$

where x and y are differentiable functions of t and the t values for the initial and endpoint are t_A and t_B, then the definite complex integral can be calculated with two real integrals. Separating the real and the imaginary parts of the integrand the integral is

$$(C)\int_A^B f(z)\,dz = \int_A^B (u\,dx - v\,dy) + \mathrm{i}\int_A^B (v\,dx + u\,dy)$$

$$= \int_{t_A}^{t_B} [u(t)x'(t) - v(t)y'(t)]\,dt + \mathrm{i}\int_{t_A}^{t_B} [v(t)x'(t) + u(t)y'(t)]\,dt \tag{14.38a}$$

with $\quad f(z) = u(x,y) + \mathrm{i}v(x,y), \quad z = x + \mathrm{i}\,y.$ (14.38b)

The notation $(C)\int_A^B f(z)\,dz$ means that the definite integral is calculated along the curve C between the points A and B. The notation $\int_{\overset{\frown}{AB}} f(z)\,dz$ is also often used, and has the same meaning.

■ $I = \int_{(C)} (z - z_0)^n\,dz \quad (n \in \mathbb{Z})$. Let the curve C be a circle around the point z_0 with radius r_0: $x = x_0 + r_0\cos t,\ y = y_0 + r_0\sin t$ with $0 \le t \le 2\pi$. For every point z of the curve C: $z = x + \mathrm{i}\,y = z_0 + r_0(\cos t + \mathrm{i}\sin t),\ dz = r_0(-\sin t + \mathrm{i}\cos t)\,dt$. After substituting these values and transforming according to the de Moivre formula follows: $I = r_0^{n+1}\int_0^{2\pi}(\cos nt + \mathrm{i}\sin nt)(-\sin t + \mathrm{i}\cos t)\,dt$ $= r_0^{n+1}\int_0^{2\pi}[\mathrm{i}\cos(n+1)t - \sin(n+1)t]\,dt = \begin{cases} 0 & \text{for } n \ne -1, \\ 2\pi\mathrm{i} & \text{for } n = -1. \end{cases}$

4. Independence of the Path of Integration

Suppose a function $f(z)$ of a complex variable is defined in a simply connected domain. The integral (14.33c) of the function can be independent of the path of integration, which connects the fixed points $A(z_A)$ and $B(z_B)$. A sufficient and necessary condition is that the function $f(z)$ is analytic in this domain, i.e., u and v satisfy the Cauchy–Riemann differential equations (14.4, p. 732). Then also the equality (14.35) is valid. If a domain is bounded by a simple closed curve, then the domain is *simply connected.*

5. Complex Integral along a Closed Curve

Suppose $f(z)$ is analytic in a simply connected domain G. Integrating the function $f(z)$ along a closed curve C which is the boundary of this domain, the value of the integral according to the Cauchy integral theorem is equal to zero (see 14.2.2, p. 747):

$$(C)\oint f(z)\,dz = 0. \tag{14.39}$$

If $f(z)$ has singular points in this domain, then the integral is calculated by using the residue theorem (see 14.3.5.5, p. 754), or by the formula (14.38a).

■ The function $f(z) = \dfrac{1}{z-a}$, with a singular point at $z = a$ has an integral value for the closed curve C directed counterclockwise around a **(Fig.14.34)** (C)$\oint \dfrac{dz}{z-a} = 2\pi\mathrm{i}\ \mathrm{Res} f(z)|_{z=a} = 2\pi\mathrm{i}.$

14.2.2 Cauchy Integral Theorem

14.2.2.1 Cauchy Integral Theorem for Simply Connected Domains

If a function $f(z)$ is analytic in a simply connected domain G, then two equivalent statements hold:

a) The integral is equal to zero along any closed curve C inside of G:

$$(C) \oint f(z)\,dz = 0. \tag{14.40}$$

b) The value of the integral $\int_A^B f(z)\,dz$ is independent of the curve connecting the points A and B and running inside of G, i.e., it depends only on A and B.
This is the *Cauchy integral theorem*.

14.2.2.2 Cauchy Integral Theorem for Multiply Connected Domains

If $C, C_1, C_2, \ldots, C_n$ are simple closed curves such that the curve C encloses all the C_ν ($\nu = 1, 2, \ldots, n$), none of the curves C_ν encloses or intersects another one, and furthermore the function $f(z)$ is analytic in a domain G which contains the curves and the region between C and C_ν, i.e., at least the region shaded in **Fig.14.35**, then

$$\oint_{(C)} f(z)\,dz = \oint_{(C_1)} f(z)\,dz + \oint_{(C_2)} f(z)\,dz + \ldots + \oint_{(C_n)} f(z)\,dz, \tag{14.41}$$

if the curves $C, C_1, \ldots, C_n$ are oriented in the same direction, e.g., counterclockwise.
This theorem is useful for the calculation of an integral along a closed curve C, if it also encloses singular points of the function $f(z)$ (see 14.3.5.5, p. 754).

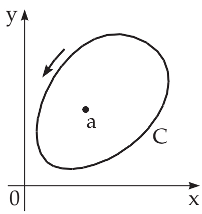


Figure 14.34

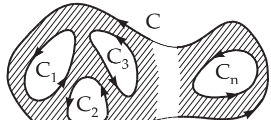


Figure 14.35

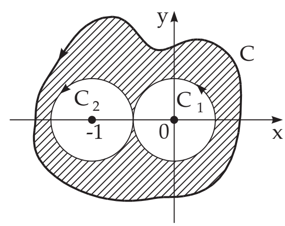


Figure 14.36

■ Calculate the integral $\oint_{(C)} \frac{z-1}{z(z+1)}\,dz$, where C is a curve enclosing the origin and the point $z = -1$ **(Fig. 14.36)**. Applying the Cauchy integral theorem, the integral along C is equal to the sum of the integrals along C_1 and C_2, where C_1 is a circle around the origin with radius $r_1 = 1/2$ and C_2 is a circle around the point $z = -1$ with radius $r_2 = 1/2$. The integrand can be decomposed into partial fractions.
Then follows: $\oint_{(C)} \frac{z-1}{z(z+1)}\,dz = \oint_{(C_1)} \frac{2\,dz}{z+1} + \oint_{(C_2)} \frac{2\,dz}{z+1} - \oint_{(C_1)} \frac{dz}{z} - \oint_{(C_2)} \frac{dz}{z} = 0 + 4\pi\mathrm{i} - 2\pi\mathrm{i} - 0 = 2\pi\mathrm{i}.$
(Compare the integrals with the example in 14.2.1.2, **3.**, p. 747.)

14.2.3 Cauchy Integral Formulas

14.2.3.1 Analytic Function on the Interior of a Domain

If $f(z)$ is analytic on a simple closed curve C and on the simply connected domain inside it, then the following representation is valid for every interior point z of this domain **(Fig. 14.37)**:

$$f(z) = \frac{1}{2\pi\mathrm{i}} \oint_{(C)} \frac{f(\zeta)}{\zeta - z}\,d\zeta \qquad (\textit{Cauchy integral formula}), \tag{14.42}$$

where ζ traces the curve C counterclockwise. With this formula, the values of an analytic function in the interior of a domain are expressed by the values of the function on the boundary of this domain. The existence and the integral representation of the n-th derivative of the function analytic on the domain G follows from (14.42):

$$f^{(n)}(z) = \frac{n!}{2\pi \mathrm{i}} \oint_{(C)} \frac{f(\zeta)}{(\zeta - z)^{n+1}} \, d\zeta. \quad (14.43)$$

Consequently, if a complex function is differentiable, i.e., it is analytic, then it is differentiable arbitrarily many times. In contrast to this, in the real case differentiability does not include repeated differentiability.
The equations (14.42) and (14.43) are called the *Cauchy integral formulas.*

14.2.3.2 Analytic Function on the Exterior of a Domain

If a function $f(z)$ is analytic on the entire part of the plane outside of a closed curve of integration C, then the values and the derivatives of the function $f(z)$ at a point z of this domain can be given with the same Cauchy formulas (14.42), (14.43), but the orientation of the curve C is now clockwise **(Fig. 14.38)**.
Also certain real integrals can be calculated with the help of the Cauchy integral formulas (see 14.4, p. 754).

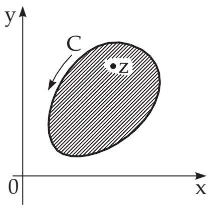


Figure 14.37

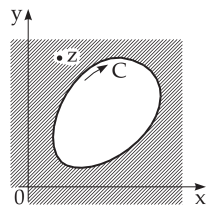


Figure 14.38

14.3 Power Series Expansion of Analytic Functions

14.3.1 Convergence of Series with Complex Terms

14.3.1.1 Convergence of a Number Sequence with Complex Terms

An infinite sequence of complex numbers $z_1, z_2, \ldots, z_n, \ldots$ has a limit z $\left(z = \lim_{n\to\infty} z_n\right)$ if for every arbitrarily given positive ε there exists an n_0 such that the inequality $|z - z_n| < \varepsilon$ holds for every $n > n_0$, i.e., from a certain n_0 the points representing the numbers $z_n, z_{n+1}, \ldots$ are inside of a circle with radius ε and center at z.

■ If the expression $\{\sqrt[n]{a}\}$ means the root with the smallest non-negative argument, then the limit $\lim_{n\to\infty} \{\sqrt[n]{a}\} = 1$ is valid for arbitrarily complex $a \neq 0$ **(Fig. 14.39)**.

14.3.1.2 Convergence of an Infinite Series with Complex Terms

A series $a_1 + a_2 + \cdots + a_n + \cdots$ with complex terms a_i converges to the number s, the sum of the series, if the sequence of the partial sums s_n with

$$s_n = a_1 + a_2 + \cdots + a_n \quad (n = 1, 2, \ldots) \quad (14.44)$$

converges. Connecting the points corresponding to the numbers $s_n = a_1 + a_2 + \cdots + a_n$ in the z plane by a broken line, then convergence means that the end of the broken line approaches the point s.

■ **A:** $\mathrm{i} + \frac{\mathrm{i}^2}{2} + \frac{\mathrm{i}^3}{3} + \frac{\mathrm{i}^4}{4} + \cdots.$ ■ **B:** $\mathrm{i} + \frac{\mathrm{i}^2}{2} + \frac{\mathrm{i}^3}{2^2} + \cdots$ **(Fig. 14.40)**.

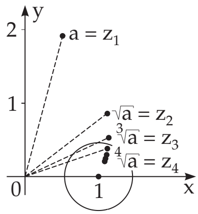


Figure 14.39

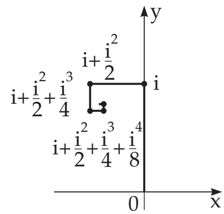


Figure 14.40

A series is called *absolutely convergent* (see ■ **B**), if the series of absolute values of the terms $|a_1| + |a_2| + |a_3| + \cdots$ is also convergent. The series is called *conditionally convergent* (see ■ **A**) if the series is convergent but not absolutely convergent. If the terms of a series are functions $f_i(z)$, like

$$f_1(z) + f_2(z) + \cdots + f_n(z) + \cdots , \tag{14.45}$$

then its sum is a function defined for the values z for which the series of the function values is convergent.

14.3.1.3 Power Series with Complex Terms

1. Convergence

A power series with complex coefficients has the form

$$P(z - z_0) = a_0 + a_1(z - z_0) + a_2(z - z_0)^2 + \cdots + a_n(z - z_0)^n + \cdots , \tag{14.46a}$$

where z_0 is a fixed point in the complex plane and the coefficients a_ν are complex constants (which can also have real values). For $z_0 = 0$ the power series has the form

$$P(z) = a_0 + a_1 z + a_2 z^2 + \cdots + a_n z^n + \cdots . \tag{14.46b}$$

If the power series $P(z - z_0)$ is convergent for a value z_1, then it is absolutely and uniformly convergent for every z in the interior of the circle with radius $r = |z_1 - z_0|$ and center at z_0.

2. Circle of Convergence

The limit between the domain of convergence and the domain of divergence of a complex power series is a uniquely defined circle. One determines its radius just as in the real case, if the limits

$$r = \frac{1}{\lim\limits_{n\to\infty} \sqrt[n]{|a_n|}} \quad \text{or} \quad r = \lim_{n\to\infty} \left| \frac{a_n}{a_{n+1}} \right| \tag{14.47}$$

exist. If the series is divergent everywhere except at $z = z_0$, then $r = 0$; if it is convergent everywhere, then $r = \infty$. The behavior of the power series on the boundary circle of the domain of convergence should be investigated point by point.

■ The power series $P(z) = \sum\limits_{n=1}^{\infty} \frac{z^n}{n}$ with radius of convergence $r = 1$ is divergent for $z = 1$ (harmonic series) and convergent for $z = -1$ (according to the Leibniz criteria for alternating series (see 7.2.3.3,**1.**, p. 463)). This power series is convergent for all further points of the unit circle $|z| = 1$ except the point $z = 1$.

3. Derivative of Power Series in the Circle of Convergence

Every power series represents an analytic function $f(z)$ inside of the circle of convergence. One gets the derivative by a term-by-term differentiation. The derivative series has the same radius of convergence as the original one.

4. Integral of Power Series in the Circle of Convergence

The power series expansion of the integral $\int_{z_0}^{z} f(\zeta)\,d\zeta$ can be obtained by a term-by-term integration of the power series of $f(z)$. The radius of convergence remains the same.

14.3.2 Taylor Series

Every function $f(z)$ analytic in a domain G can be expanded uniquely into a power series of the form

$$f(z) = \sum_{n=0}^{\infty} a_n (z - z_0)^n \qquad \text{(Taylor series)}, \tag{14.48a}$$

for any z_0 in G, where the circle of convergence is the greatest circle around z_0 which belongs entirely to the domain G **(Fig. 14.41)**. The coefficients a_n are complex numbers in general; for them one gets:

$$a_n = \frac{f^{(n)}(z_0)}{n!}\,. \tag{14.48b}$$

The Taylor series can be written in the form

$$f(z) = f(z_0) + \frac{f'(z_0)}{1!}(z - z_0) + \frac{f''(z_0)}{2!}(z - z_0)^2 + \cdots + \frac{f^{(n)}(z_0)}{n!}(z - z_0)^n + \cdots\,. \tag{14.48c}$$

Every power series is the Taylor expansion of its sum function in the interior of its circle of convergence.

■ Examples of Taylor expansions are the series representations of the functions e^z, $\sin z$, $\cos z$, $\sinh z$, and $\cosh z$ in 14.5.2, p. 758.

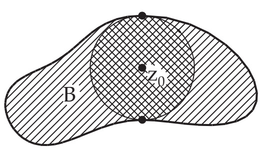


Figure 14.41

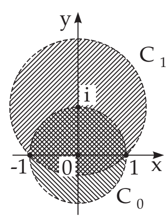


Figure 14.42

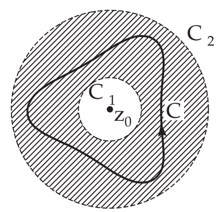


Figure 14.43

14.3.3 Principle of Analytic Continuation

Let consider the case when the circles of convergence K_0 around z_0 and K_1 around z_1 of the two power series

$$f_0(z) = \sum_{n=0}^{\infty} a_n (z - z_0)^n \qquad \text{and} \qquad f_1(z) = \sum_{n=0}^{\infty} b_n (z - z_1)^n \tag{14.49a}$$

have a certain common domain **(Fig. 14.42)** and in this domain they are equal:

$$f_0(z) = f_1(z). \tag{14.49b}$$

Then both power series are the Taylor expansions of the same analytic function $f(z)$, belonging to the points z_0 and z_1. The function $f_1(z)$ is called the *analytic continuation* into K_1 of the function $f_0(z)$ defined only in K_0.

■ The geometric series $f_0(z) = \sum_{n=0}^{\infty} z^n$ with the circle of convergence K_0 $(r_0 = 1)$ around $z_0 = 0$ and $f_1(z) = \frac{1}{1-\mathrm{i}} \sum_{n=0}^{\infty} \left(\frac{z - \mathrm{i}}{1 - \mathrm{i}}\right)^n$ with the circle of convergence K_1 $(r_1 = \sqrt{2})$ around $z_1 = \mathrm{i}$ have the same

analytic function $f(z) = 1/(1-z)$ as their sum in their own circle of convergence, consequently also on the common part of them (doubly shaded region in **Fig. 14.42**) for $z \neq 1$. So, $f_1(z)$ is the analytic continuation of $f_0(z)$ from K_0 into K_1 (and conversely).

14.3.4 Laurent Expansion

Every function $f(z)$, which is analytic in the interior of a circular ring between two concentric circles with center z_0 and radii r_1 and r_2, can be expanded into a generalized power series, into the so-called Laurent series:

$$f(z) = \sum_{n=-\infty}^{\infty} a_n(z-z_0)^n = \cdots + \frac{a_{-k}}{(z-z_0)^k} + \frac{a_{-k+1}}{(z-z_0)^{k-1}} + \cdots$$
$$+ \frac{a_{-1}}{z-z_0} + a_0 + a_1(z-z_0) + a_2(z-z_0)^2 + \cdots + a_k(z-z_0)^k + \cdots . \quad (14.50a)$$

The coefficients a_n are usually complex and they are uniquely defined by the formula

$$a_n = \frac{1}{2\pi i} \oint_{(C)} \frac{f(\zeta)}{(\zeta - z_0)^{n+1}}\, d\zeta \qquad (n = 0, \pm 1, \pm 2, \ldots), \quad (14.50b)$$

where C denotes an arbitrary closed curve which is in the circular ring $r_1 < |z| < r_2$, and the circle with radius r_1 is inside of it, and its orientation is counterclockwise **(Fig. 14.43)**. If the domain G of the function $f(z)$ is larger than the circular ring, then the domain of convergence of the *Laurent series* is the largest circular ring with center z_0 lying entirely in G.

■ The Laurent series expansion of the function $f(z) = \dfrac{1}{(z-1)(z-2)}$ is determined around $z_0 = 0$ in the circular ring $1 < |z| < 2$ where $f(z)$ is analytic. First the function $f(z)$ is decomposed into partial fractions: $f(z) = \dfrac{1}{z-2} - \dfrac{1}{z-1}$. Since $|1/z| < 1$ and $|z/2| < 1$ holds in the considered domain, the two terms of this decomposition can be written as the sums of geometric series being absolutely convergent in the entire circular ring $1 < |z| < 2$. One gets:

$$f(z) = \frac{1}{(z-1)(z-2)} = -\frac{1}{z\left(1-\dfrac{1}{z}\right)} - \frac{1}{2\left(1-\dfrac{z}{2}\right)} = -\underbrace{\sum_{n=1}^{\infty}\frac{1}{z^n}}_{|z|>1} - \frac{1}{2}\underbrace{\sum_{n=0}^{\infty}\left(\frac{z}{2}\right)^n}_{|z|<2}$$
$$= \sum_{n=-\infty}^{\infty} a_n z^n \quad \text{with} \quad a_n = \begin{cases} -1 & \text{for } n = -1, -2, \ldots, \\ \dfrac{-1}{2^{n+1}} & \text{for } n = 0, 1, 2, \ldots . \end{cases}$$

14.3.5 Isolated Singular Points and the Residue Theorem

14.3.5.1 Isolated Singular Points

If a function $f(z)$ is analytic in the neighborhood of a point z_0 but not at the point z_0 itself, then z_0 is called an *isolated singular point* of the function $f(z)$. If $f(z)$ can be expanded into a Laurent series in the neighborhood of z_0

$$f(z) = \sum_{n=-\infty}^{\infty} a_n(z-z_0)^n, \quad (14.51)$$

then the isolated singular point can be classified by the behavior of the Laurent series:

1. If the Laurent series does not contain any term with a negative power of $(z-z_0)$, i.e., $a_n = 0$ for

$n < 0$ holds, then the Laurent series is a Taylor series with coefficients given by the Cauchy integral formula

$$a_n = \frac{1}{2\pi \mathrm{i}} \oint\limits_{(K)} (\zeta - z_0)^{-n-1} f(\zeta)\, d\zeta = \frac{f^{(n)}(z_0)}{n!} \quad (n = 0, 1, 2, \ldots). \tag{14.52}$$

In this case, the function $f(z)$ itself is either analytic at the point z_0 and $f(z_0) = a_0$ or z_0 is a removable singularity.

2. If the Laurent series contains a finite number of terms with negative powers of $(z - z_0)$, i.e., $a_m \neq 0$, $a_n = 0$ for $n < m < 0$, then z_0 is called a *pole*, a *pole of order* m, or a *pole of multiplicity* m. Multiplying by $(z - z_0)^m$, and not by any lower power, then $f(z)$ is transformed into a function which is analytic at z_0 and in its neighborhood.

■ $f(z) = \frac{1}{2}\left(z + \frac{1}{z}\right)$ has a pole of order one at $z = 0$.

3. If the Laurent series contains an infinite number of terms with negative powers of $(z - z_0)$, then z_0 is an *essential singularity* of the function $f(z)$.
Approaching a pole, $|f(z)|$ tends to ∞. Approaching an essential singularity, $f(z)$ gets arbitrarily close to any complex number c.

■ The function $f(z) = e^{1/z}$, whose Laurent series is $f(z) = \sum_{n=0}^{\infty} \frac{1}{n!}\frac{1}{z^n}$, has an essential singularity at $z = 0$.

14.3.5.2 Meromorphic Functions

If an otherwise holomorphic function has only isolated singularities which are poles, then it is called *meromorphic*. A meromorphic function can always be represented as the quotient of two analytic functions.

■ Examples of functions meromorphic on the whole plane are the rational functions which have a finite number of poles, and also transcendental functions such as $\tan z = \frac{\sin z}{\cos z}$ and $\cot z = \frac{\cos z}{\sin z}$.

14.3.5.3 Elliptic Functions

Elliptic functions are double periodic functions whose singularities are poles, i.e., they are meromorphic functions with two independent periods (see 14.6, p. 762). If the two periods are ω_1 and ω_2, which are in a non-real relation, then

$$f(z + m\omega_1 + n\omega_2) = f(z) \quad \left(m, n = 0, \pm 1, \pm 2, \ldots; \operatorname{Im}\left(\frac{\omega_1}{\omega_2}\right) \neq 0\right). \tag{14.53}$$

The range of $f(z)$ is already attained in a primitive period parallelogram with the points $0, \omega_1, \omega_1 + \omega_2, \omega_2$.

14.3.5.4 Residue

With z_0 as an isolated singularity of the function $f(z)$ the coefficient a_{-1} of the power $(z - z_0)^{-1}$ in the Laurent expansion of $f(z)$ valid in the neighborhood of z_0 is called the *residue of the function* $f(z)$ at the point z_0. According to (14.50b)

$$a_{-1} = \operatorname{Res} f(z)|_{z=z_0} = \frac{1}{2\pi \mathrm{i}} \oint\limits_{(K)} f(\zeta)\, d\zeta \tag{14.54a}$$

holds. The residue belonging to a pole of order m can be calculated by the formula

$$a_{-1} = \operatorname{Res} f(z)|_{z=z_0} = \lim_{z \to z_0} \frac{1}{(m-1)!} \frac{d^{m-1}}{dz^{m-1}} [f(z)(z - z_0)^m]. \tag{14.54b}$$

If the function $f(z)$ can be represented as a quotient $f(z) = \varphi(z)/\psi(z)$, where the functions $\varphi(z)$ and $\psi(z)$ are analytic at the point $z = z_0$ and z_0 is a simple root of the function $\psi(z)$, i.e., $\psi(z_0) = 0$ and $\psi'(z_0) \neq 0$ holds, then the point $z = z_0$ is a pole of order one of the function $f(z)$. It follows from (14.54b) that

$$\operatorname{Res}\left[\frac{\varphi(z)}{\psi(z)}\right]_{z=z_0} = \frac{\varphi(z_0)}{\psi'(z_0)}. \tag{14.54c}$$

If z_0 is a root of multiplicity m of the function $\psi(z)$, i.e., $\psi(z_0) = \psi'(z_0) = \cdots = \psi^{(m-1)}(z_0) = 0$, $\psi^{(m)}(z_0) \neq 0$ holds, then the point $z = z_0$ is a pole of order m of $f(z)$.

14.3.5.5 Residue Theorem

With the help of residues one can calculate the integral of a function along a closed curve enclosing isolated singular points **(Fig. 14.44)**.
If the function $f(z)$ is single valued and analytic in a simply connected domain G except at a finite number of points $z_0, z_1, z_2, \ldots, z_n$, and the domain is bounded by the closed curve C, then the value of the integral of the function along this closed curve in a counterclockwise direction is the product of $2\pi\mathrm{i}$ and the sum of the residues in all these singular points:

$$\oint_{(K)} f(z)\,dz = 2\pi\mathrm{i}\sum_{k=0}^{n} \operatorname{Res} f(z)\,|_{z=z_k}. \tag{14.55}$$

■ The function $f(z) = e^z/(z^2+1)$ has poles of order one at $z_{1,2} = \pm\mathrm{i}$. The corresponding residues have the sum $\sin 1$. If K is a circle around the origin with radius $r > 1$, then

$$\oint_{(K)} \frac{e^z}{z^2+1}\,dz = 2\pi\mathrm{i}\left(\frac{e^{z_1}}{2z_1} + \frac{e^{z_2}}{2z_2}\right) = 2\pi\mathrm{i}\left(\frac{e^{\mathrm{i}}}{2\mathrm{i}} - \frac{e^{-\mathrm{i}}}{2\mathrm{i}}\right) = 2\pi\mathrm{i}\sin 1.$$

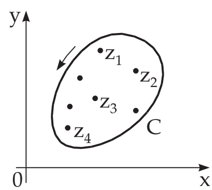


Figure 14.44

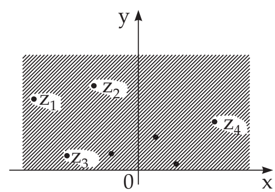


Figure 14.45

14.4 Evaluation of Real Integrals by Complex Integrals

14.4.1 Application of Cauchy Integral Formulas

The value of certain real integrals can be calculated with the help of the Cauchy integral formula.

■ The function $f(z) = e^z$, which is analytic in the whole z plane, can be represented with the Cauchy integral formula (14.42), where the path of integration C is a circle with center z and radius r. The equation of the circle is $\zeta = z + re^{\mathrm{i}\varphi}$. From (14.43) follows

$$e^z = \frac{n!}{2\pi\mathrm{i}}\oint_{(C)} \frac{e^\zeta}{(\zeta - z)^{n+1}}\,d\zeta = \frac{n!}{2\pi\mathrm{i}}\int_{\varphi=0}^{\varphi=2\pi} \frac{e^{(z+re^{\mathrm{i}\varphi})}}{r^{n+1}e^{\mathrm{i}\varphi(n+1)}}\mathrm{i}re^{\mathrm{i}\varphi}\,d\varphi = \frac{n!}{2\pi r^n}\int_0^{2\pi} e^{z+r\cos\varphi+\mathrm{i}r\sin\varphi-\mathrm{i}n\varphi}\,d\varphi, \text{ so that}$$

$$\frac{2\pi r^n}{n!} = \int_0^{2\pi} e^{r\cos\varphi+\mathrm{i}(r\sin\varphi-n\varphi)}\,d\varphi = \int_0^{2\pi} e^{r\cos\varphi}[\cos(r\sin\varphi - n\varphi)]\,d\varphi + \mathrm{i}\int_0^{2\pi} e^{r\cos\varphi}[\sin(r\sin\varphi - n\varphi)]\,d\varphi.$$

Comparing real and imaginary part gives $\int_0^{2\pi} e^{r\cos\varphi}\cos(r\sin\varphi - n\varphi)\,d\varphi = \frac{2\pi r^n}{n!}$, since the imaginary part is equal to zero.

14.4.2 Application of the Residue Theorem

Several definite integrals of real functions with one variable can be calculated with the help of the residue theorem. If $f(z)$ is a function which is analytic in the whole upper half of the complex plane including the real axis except the singular points $z_1, z_2, \ldots, z_n$ above the real axis **(Fig. 14.45)**, and if one of the roots of the equation $f(1/z) = 0$ has multiplicity $m \geq 2$ (see 1.6.3.1, **1.**, p. 43), then

$$\int_{-\infty}^{+\infty} f(x)\,dx = 2\pi\mathrm{i}\sum_{i=1}^{n} \operatorname{Res} f(z)|_{z=z_i}. \tag{14.56}$$

■ Calculation of the integral $\int_{-\infty}^{+\infty} \frac{dx}{(1+x^2)^3}$: The equation $f\left(\frac{1}{x}\right) = \frac{1}{\left(1+\frac{1}{x^2}\right)^3} = \frac{x^6}{(x^2+1)^3} = 0$ has a root of order six at $x = 0$. The function $w = \frac{1}{(1+z^2)^3}$ has a single singular point $z = \mathrm{i}$ in the upper half-plane, which is a pole of order 3, since the equation $(1+z^2)^3 = 0$ has two triple roots at i and $-$i. The residue is according to (14.54b):
$\operatorname{Res}\frac{1}{(1+z^2)^3|_{z=\mathrm{i}}} = \frac{1}{2!}\frac{d^2}{dz^2}\left[\frac{(z-\mathrm{i})^3}{(1+z^2)^3}\right]_{z=\mathrm{i}}$. From $\frac{d^2}{dz^2}\left(\frac{z-\mathrm{i}}{1+z^2}\right)^3 = \frac{d^2}{dz^2}(z+\mathrm{i})^{-3} = 12(z+\mathrm{i})^{-5}$ it follows that $\operatorname{Res}\frac{1}{(1+z^2)^3|_{z=\mathrm{i}}} = 6(z+\mathrm{i})^{-5}|_{z=\mathrm{i}} = \frac{6}{(2\mathrm{i})^5} = -\frac{3}{16}\mathrm{i}$, and with (14.56): $\int_{-\infty}^{+\infty} f(x)\,dx = 2\pi\mathrm{i}\left(-\frac{3}{16}\mathrm{i}\right) = \frac{3}{8}\pi$. For further applications of residue theory see, e.g., [14.12].

14.4.3 Application of the Jordan Lemma

14.4.3.1 Jordan Lemma

In many cases, real improper integrals with an infinite interval of integration can be calculated by complex integrals along a closed curve. To avoid the always recurrent estimations, the *Jordan lemma* is used about improper integrals of the form

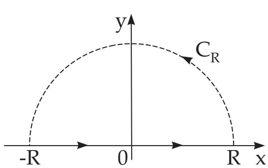


Figure 14.46

$$\int_{(C_R)} f(z)e^{\mathrm{i}\alpha z}\,dz, \tag{14.57a}$$

where C_R is the half-circle arc with center at the origin and with the radius R in the upper half of the z plane **(Fig. 14.46)**. The Jordan lemma distinguishes the following cases:

a) $\boldsymbol{\alpha > 0}$**:** If $f(z)$ tends to zero uniformly in the upper half-plane and also on the real axis for $|z| \to \infty$ and if $\alpha > 0$ holds, then for $R \to \infty$ is valid:

$$\int_{(C_R)} f(z)e^{\mathrm{i}\alpha z}\,dz \to 0\,. \tag{14.57b}$$

b) $\boldsymbol{\alpha = 0}$**:** If the expression $z\,f(z)$ tends to zero uniformly for $|z| \to \infty$, then the above statement (14.57b) is also valid in the case $\alpha = 0$.

c) $\boldsymbol{\alpha < 0}$**:** If the half-circle C_R is now below the real axis, then the statement corresponding to (14.57b) is also valid for $\alpha < 0$.

d) The statement to (14.57b) is also valid if only an arc segment is considered instead of the complete half-circle.

e) When C_R^* is a half-circle or an arc segment in the left half-plane with $\alpha > 0$, or in the right one with $\alpha < 0$, then the statement corresponding to (14.57b) is valid for the integral in the form

$$\int\limits_{(C_R^*)} f(z)e^{\alpha z}\,dz\,. \tag{14.57c}$$

14.4.3.2 Examples of the Jordan Lemma

1. Evaluation of the Integral $\displaystyle\int\limits_0^\infty \frac{x\sin\alpha x}{x^2+a^2}\,dx \quad (\alpha > 0, a \ge 0)\,.$ (14.58a)

The following complex integral is assigned to the above real integral:

$$2\mathrm{i}\underbrace{\int\limits_0^R \frac{x\sin\alpha x}{x^2+a^2}}_{\text{even function}}dx = \mathrm{i}\int\limits_{-R}^R \frac{x\sin\alpha x}{x^2+a^2}\,dx + \underbrace{\int\limits_{-R}^R \frac{x\cos\alpha x}{x^2+a^2}\,dx}_{=0\ \text{(odd integrand)}} = \int\limits_{-R}^R \frac{xe^{\mathrm{i}\alpha x}}{x^2+a^2}\,dx. \tag{14.58b}$$

The very last of these integrals is part of the complex integral $\displaystyle\oint\limits_{(C)} \frac{ze^{\mathrm{i}\alpha z}}{z^2+a^2}\,dz$. The curve C contains the C_R half-circle defined above and the part of the real axis between the values $-R$ and R $(R > |a|)$. The complex integrand has the only singular point in the upper half-plane $z = a\,\mathrm{i}$. From the residue theorem follows: $\displaystyle I = \oint\limits_{(C)} \frac{ze^{\mathrm{i}\alpha z}}{z^2+a^2}\,dz = 2\pi\mathrm{i}\lim_{z\to a\mathrm{i}}\left[\frac{ze^{\mathrm{i}\alpha z}}{z^2+a^2}(z-a\mathrm{i})\right] = 2\pi\mathrm{i}\lim_{z\to a\mathrm{i}}\frac{ze^{\mathrm{i}}\alpha z}{z+a\mathrm{i}} = \pi\mathrm{i}e^{-\alpha a}$, hence

$\displaystyle I = \int\limits_{(C_R)} \frac{ze^{\mathrm{i}\alpha z}}{z^2+a^2}\,dz + \int\limits_{-R}^R \frac{xe^{\mathrm{i}\alpha x}}{x^2+a^2}\,dx = \pi\mathrm{i}e^{-\alpha a}$. From $\lim\limits_{R\to\infty} I$ and from the Jordan lemma follows

$$\int\limits_0^\infty \frac{x\sin\alpha x}{x^2+a^2}\,dx = \frac{\pi}{2}e^{-\alpha a} \quad (\alpha > 0,\ \ a \ge 0). \tag{14.58c}$$

Several further integrals can be evaluated in a similar way (see **Table 21.8**, p. 1098).

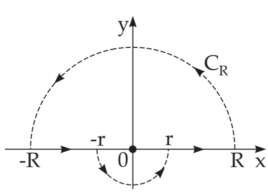


Figure 14.47

2. Sine Integral (see also 8.2.5,**1.**, (8.95), p. 513)

The integral $\displaystyle\int\limits_0^\infty \frac{\sin x}{x}\,dx$ is called the *sine integral* or the *integral sine* (see also 8.2.5, **1.**, p. 513). Analogously to the previous example, here the complex integral $\displaystyle I = \oint\limits_{C_R} \frac{e^{\mathrm{i}z}}{z}\,dz$ is investigated with the curve C_R according to **Fig. 14.47**. The integrand of the complex integral has a pole of first order at $z = 0$, so

$\displaystyle I = 2\pi\mathrm{i}\lim_{z\to 0}\left[\frac{e^{\mathrm{i}z}}{z}z\right] = 2\pi\mathrm{i}$, hence $\displaystyle I = 2\mathrm{i}\int\limits_r^R \frac{\sin x}{x}\,dx + \mathrm{i}\int\limits_\pi^{2\pi} e^{\mathrm{i}r(\cos\varphi + \mathrm{i}\sin\varphi)}\,d\varphi + \int\limits_{C_R} \frac{e^{\mathrm{i}z}}{z}\,dz = 2\pi\mathrm{i}$. This limit is evaluated as $R \to \infty$, $r \to 0$, where the second integral tends to 1 uniformly for $r \to 0$ with

respect to φ, i.e., the limiting process $r \to 0$ can be done behind the integral sign. With the *Jordan lemma* follows:

$$2\mathrm{i}\int\limits_0^\infty \frac{\sin x}{x}\,dx + \pi\mathrm{i} = 2\pi\mathrm{i}, \quad \text{hence} \quad \int\limits_0^\infty \frac{\sin x}{x}\,dx = \frac{\pi}{2}\,. \tag{14.59}$$

3. Step Function

Discontinuous real functions can be represented as complex integrals. The so-called *step function* (see also 15.2.1.3, p. 774) is an example:

$$F(t) = \frac{1}{2\pi\mathrm{i}}\int\limits_{-\smile\to} \frac{e^{\mathrm{i}tz}}{z}\,dz = \begin{cases} 1 & \text{for } t > 0\,,\\ 1/2 & \text{for } t = 0\,,\\ 0 & \text{for } t < 0\,.\end{cases} \tag{14.60}$$

The symbol $-\smile\to$ denotes a path of integration along the real axis ($|\,R\,| \to \infty$) going round the origin **(Fig. 14.47)**.
If t denotes time, then the function $\Phi(t) = cF(t - t_0)$ represents a quantity which jumps at time $t = t_0$ from 0 through the value $c/2$ to the value c. It is called a step function or also a *Heaviside function*. It is used in the electrotechnics to describe suddenly occurring voltage or current jumps.

4. Rectangular Pulse

A further example of the application of complex integrals and the Jordan lemma is the representation of the rectangular pulse (see also 15.2.1.3, p. 774):

$$\Psi(t) = \frac{1}{2\pi\mathrm{i}}\int\limits_{-\smile\to} \frac{e^{\mathrm{i}(b-t)z}}{z}\,dz - \frac{1}{2\pi\mathrm{i}}\int\limits_{-\smile\to} \frac{e^{\mathrm{i}(a-t)z}}{z}\,dz = \begin{cases} 0 & \text{for } t < a \text{ and } t > b,\\ 1 & \text{for } a < t < b,\\ 1/2 & \text{for } t = a \text{ and } t = b.\end{cases} \tag{14.61}$$

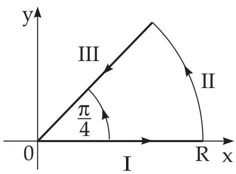


Figure 14.48

5. Fresnel Integrals

To derive the *Fresnel integral*

$$\int\limits_0^\infty \sin(x^2)\,dx = \int\limits_0^\infty \cos(x^2)\,dx = \frac{1}{2}\sqrt{\pi/2} \tag{14.62}$$

the integral $I = \int\limits_K e^{-z^2}\,dz$ has to be investigated on the closed path of integration shown in **Fig. 14.48**. According to the Cauchy integral theorem holds: $I = I_{\mathrm{I}} + I_{\mathrm{II}} + I_{\mathrm{III}} = 0$ with $I_{\mathrm{I}} = \int_0^R e^{-x^2}\,dx\,,\ I_{\mathrm{II}} = \mathrm{i}R\int_0^{\pi/4} e^{-R^2(\cos 2\varphi + \mathrm{i}\sin 2\varphi)+\mathrm{i}\varphi}d\varphi,$ $I_{\mathrm{III}} = e^{\frac{\mathrm{i}\pi}{4}}\int_R^0 e^{\mathrm{i}r^2}dr = \frac{1}{2}\sqrt{2}(1+\mathrm{i})\left[\mathrm{i}\int_0^R \sin r^2\,dr - \int_0^R \cos r^2\,dr\right]$. Estimation of I_{II}: Since $|\mathrm{i}| = |e^{\mathrm{i}\tau}| = 1$ (τ real) holds, one gets: $|I_{\mathrm{II}}| \le R\int_0^{\pi/4} e^{-R^2\cos 2\varphi}\,d\varphi = \frac{R}{2}\int_0^{\alpha} e^{-R^2\cos\varphi}\,d\varphi + \frac{R}{2}\int_\alpha^{\pi/2} e^{-R^2\cos\varphi}\,d\varphi <$ $\frac{R}{2}\int_0^{\alpha} e^{-R^2\cos\alpha}\,d\varphi + \frac{R}{2}\int_\alpha^{\pi/2}\frac{\sin\varphi}{\sin\alpha}e^{-R^2\cos\varphi}\,d\varphi < \frac{\alpha R}{2}e^{-R^2\cos\alpha} + \frac{1 - e^{-R^2\cos\alpha}}{2R\sin\alpha}$ $\left(0 < \alpha < \frac{\pi}{2}\right)$. Performing the limiting process $\lim\limits_{R\to\infty} I$ yields the values of the integrals I_{I} and I_{II}: $\lim\limits_{R\to\infty} I_{\mathrm{I}} = \frac{1}{2}\sqrt{\pi}\,,\ \lim\limits_{R\to\infty} I_{\mathrm{II}} = 0$. The given formulas (14.62) can be obtained by separating the real and imaginary parts.

14.5 Algebraic and Elementary Transcendental Functions

14.5.1 Algebraic Functions

1. Definition

A function which is the result of finitely many algebraic operations performed with z and maybe also with finitely many constants, is called an *algebraic function*. In general, a complex algebraic function $w(z)$ can be defined in an implicit way as a polynomial, just as its real analogue

$$a_1 z^{m_1} w^{n_1} + a_2 z^{m_2} w^{n_2} + \cdots + a_k z^{m_k} w^{n_k} = 0. \tag{14.63}$$

It happens that w cannot be expressed explicitly.

2. Examples of Algebraic Functions

Linear function: $w = az + b$. (14.64)

Inverse function: $w = \dfrac{1}{z}$. (14.65)

Quadratic function: $w = z^2$. (14.66)

Square root function: $w = \sqrt{z^2 - a^2}$. (14.67)

Fractional linear function: $w = \dfrac{z + \mathrm{i}}{z - \mathrm{i}}$. (14.68)

14.5.2 Elementary Transcendental Functions

The complex transcendental functions have definitions corresponding to the transcendental real functions, just as in the case of the algebraic functions. For a detailed discussion of them see, e.g., [21.1] or [21.12].

1. Natural Exponential Function

$$e^z = 1 + \frac{z}{1!} + \frac{z^2}{2!} + \frac{z^3}{3!} + \cdots . \tag{14.69}$$

The series is absolutely convergent in the whole z plane.

a) Pure imaginary exponent iy: This is valid according to the *Euler relation* (see 1.5.2.4, p. 36):

$$e^{\mathrm{i}y} = \cos y + \mathrm{i}\sin y \quad \text{with} \quad e^{\pi\mathrm{i}} = -1. \tag{14.70}$$

b) General case $z = x + \mathrm{i}y$:

$$e^z = e^{x+\mathrm{i}y} = e^x e^{\mathrm{i}y} = e^x(\cos y + \mathrm{i}\sin y), \tag{14.71a}$$

$$\mathrm{Re}\,(e^z) = e^x \cos y, \quad \mathrm{Im}\,(e^z) = e^x \sin y, \quad |e^z| = e^x, \quad \arg(e^z) = y. \tag{14.71b}$$

The function e^z is periodic, its period is $2\pi\,\mathrm{i}$: $\quad e^z = e^{z+2k\pi\,\mathrm{i}} \quad (k = 0, \pm 1. \pm 2, \ldots)$. (14.71c)

In particular: $\quad e^0 = e^{2k\pi\,\mathrm{i}} = 1, \qquad e^{(2k+1)\pi\mathrm{i}} = -1.$ (14.71d)

c) Exponential form of a complex number (see 1.5.2.4, p. 36):

$$a + \mathrm{i}b = \rho e^{\mathrm{i}\varphi}. \tag{14.72}$$

d) *Euler relation for complex numbers*:

$$e^{\mathrm{i}z} = \cos z + \mathrm{i}\sin z, \tag{14.73a}$$

$$e^{-\mathrm{i}z} = \cos z - \mathrm{i}\sin z. \tag{14.73b}$$

2. Natural Logarithm

$$w = \mathrm{Ln}\,z, \quad \text{if} \quad z = e^w. \tag{14.74a}$$

Since $z = \rho e^{\mathrm{i}\,\varphi}$, one can write:

$$\mathrm{Ln}\,z = \ln\rho + \mathrm{i}\,(\varphi + 2k\pi) \quad \text{and} \tag{14.74b}$$

$$\mathrm{Re}\,(\mathrm{Ln}\,z) = \ln\rho, \quad \mathrm{Im}\,(\mathrm{Ln}\,z) = \varphi + 2k\,\pi \qquad (k = 0, \pm 1, \pm 2, \ldots). \tag{14.74c}$$

Since $\operatorname{Ln} z$ is a multiple-valued function (see 2.8.2, p. 86), we usually give only the *principal value of the logarithm* $\ln z$:

$$\ln z = \ln \rho + \mathrm{i}\,\varphi \qquad (-\pi < \varphi \le +\pi). \tag{14.74d}$$

The function $\operatorname{Ln} z$ is defined for every complex number, except for $z = 0$.

3. General Exponential Function

$$a^z = e^{z \operatorname{Ln} a}. \tag{14.75a}$$

a^z $(a \ne 0)$ is a multiple-valued function (see 2.8.2, p. 86) with principal value

$$a^z = e^{z \ln a}. \tag{14.75b}$$

4. Trigonometric Functions and Hyperbolic Functions

$$\sin z = \frac{e^{\mathrm{i}z} - e^{-\mathrm{i}z}}{2\mathrm{i}} = z - \frac{z^3}{3!} + \frac{z^5}{5!} - \cdots, \tag{14.76a}$$

$$\cos z = \frac{e^{\mathrm{i}z} + e^{-\mathrm{i}z}}{2} = 1 - \frac{z^2}{2!} + \frac{z^4}{4!} - \cdots, \tag{14.76b}$$

$$\sinh z = \frac{e^{z} - e^{-z}}{2} = z + \frac{z^3}{3!} + \frac{z^5}{5!} + \cdots, \tag{14.77a}$$

$$\cosh z = \frac{e^{z} + e^{-z}}{2} = 1 + \frac{z^2}{2!} + \frac{z^4}{4!} + \cdots. \tag{14.77b}$$

All four series are convergent on the entire z plane and they are all periodic. The period of the functions (14.76a,b) is 2π, the period of the functions (14.77a,b) is $2\pi\,\mathrm{i}$.
The relations between these functions for any real or complex z are:

$$\sin \mathrm{i}\,z = \mathrm{i}\sinh z, \tag{14.78a}$$ $$\cos \mathrm{i}\,z = \cosh z, \tag{14.78b}$$

$$\sinh \mathrm{i}\,z = \mathrm{i}\sin z, \tag{14.79a}$$ $$\cosh \mathrm{i}\,z = \cos z. \tag{14.79b}$$

The transformation formulas of the real trigonometric and hyperbolic functions (see 2.7.2, p. 81, and 2.9.3, p. 91) are also valid for the complex functions. The values of the functions $\sin z$, $\cos z$, $\sinh z$, and $\cosh z$ for the argument $z = x + \mathrm{i}\,y$ can be calculated with the help of the formulas $\sin(a+b)$, $\cos(a+b)$, $\sinh(a+b)$, and $\cosh(a+b)$ or by using the Euler relation (see 1.5.2.4, p. 36).

■ $\cos(x + \mathrm{i}\,y) = \cos x \cos \mathrm{i}\,y - \sin x \sin \mathrm{i}\,y = \cos x \cosh y - \mathrm{i} \sin x \sinh y.$ (14.80)

Therefore:

$$\operatorname{Re}(\cos z) = \cos \operatorname{Re}(z)\,\cosh \operatorname{Im}(z), \tag{14.81a}$$

$$\operatorname{Im}(\cos z) = -\sin \operatorname{Re}(z)\,\sinh \operatorname{Im}(z). \tag{14.81b}$$

The functions $\tan z$, $\cot z$, $\tanh z$, and $\coth z$ are defined by the following formulas:

$$\tan z = \frac{\sin z}{\cos z}, \quad \cot z = \frac{\cos z}{\sin z}, \tag{14.82a}$$ $$\tanh z = \frac{\sinh z}{\cosh z}, \quad \coth z = \frac{\cosh z}{\sinh z}. \tag{14.82b}$$

5. Inverse Trigonometric Functions and Inverse Hyperbolic Functions

These functions are many-valued functions, and one can express them with the help of the logarithm function:

$$\operatorname{Arcsin} z = -\mathrm{i} \operatorname{Ln}\left(\mathrm{i}\,z + \sqrt{1 - z^2}\right), \tag{14.83a}$$ $$\operatorname{Arsinh} z = \operatorname{Ln}\left(z + \sqrt{z^2 + 1}\right), \tag{14.83b}$$

$$\operatorname{Arccos} z = -\mathrm{i} \operatorname{Ln}\left(z + \sqrt{z^2 - 1}\right), \tag{14.84a}$$ $$\operatorname{Arcosh} z = \operatorname{Ln}\left(z + \sqrt{z^2 - 1}\right), \tag{14.84b}$$

$$\operatorname{Arctan} z = \frac{1}{2\mathrm{i}} \operatorname{Ln} \frac{1 + \mathrm{i}\,z}{1 - \mathrm{i}\,z}, \tag{14.85a}$$ $$\operatorname{Artanh} z = \frac{1}{2} \operatorname{Ln} \frac{1 + z}{1 - z}, \tag{14.85b}$$

$$\operatorname{Arccot} z = -\frac{1}{2\mathrm{i}}\operatorname{Ln}\frac{\mathrm{i}\,z+1}{\mathrm{i}\,z-1}, \quad (14.86a) \qquad \operatorname{Arcoth} z = \frac{1}{2}\operatorname{Ln}\frac{z+1}{z-1}. \quad (14.86b)$$

The *principal values* of the inverse trigonometric and the inverse hyperbolic functions can be expressed by the same formulas using the principal value of the logarithm $\ln z$:

$$\arcsin z = -\mathrm{i}\,\ln(\mathrm{i}\,z + \sqrt{1-z^2}), \quad (14.87a) \qquad \operatorname{arsinh} z = \ln(z + \sqrt{z^2+1}), \quad (14.87b)$$

$$\arccos z = -\mathrm{i}\,\ln(z + \sqrt{z^2-1}), \quad (14.88a) \qquad \operatorname{arcosh} z = \ln(z + \sqrt{z^2-1}), \quad (14.88b)$$

$$\arctan z = \frac{1}{2\mathrm{i}}\ln\frac{1+\mathrm{i}z}{1-\mathrm{i}z}, \quad (14.89a) \qquad \operatorname{artanh} z = \frac{1}{2}\ln\frac{1+z}{1-z}, \quad (14.89b)$$

$$\operatorname{arccot} z = -\frac{1}{2\mathrm{i}}\ln\frac{\mathrm{i}z+1}{\mathrm{i}z-1}, \quad (14.90a) \qquad \operatorname{arcoth} z = \frac{1}{2}\ln\frac{z+1}{z-1}. \quad (14.90b)$$

6. Real and Imaginary Part of the Trigonometric and Hyperbolic Functions (See Table 14.1)

Table 14.1 Real and imaginary parts of the trigonometric and hyperbolic functions

Function $w = f(x+\mathrm{i}y)$	**Real part Re** (w)	**Imaginary part Im** (w)
$\sin(x \pm \mathrm{i}y)$	$\sin x \cosh y$	$\pm\cos x \sinh y$
$\cos(x \pm \mathrm{i}y)$	$\cos x \cosh y$	$\mp\sin x \sinh y$
$\tan(x \pm \mathrm{i}y)$	$\dfrac{\sin 2x}{\cos 2x + \cosh 2y}$	$\pm\dfrac{\sinh 2y}{\cos 2x + \cosh 2y}$
$\sinh(x \pm \mathrm{i}y)$	$\sinh x \cos y$	$\pm\cosh x \sin y$
$\cosh(x \pm \mathrm{i}y)$	$\cosh x \cos y$	$\pm\sinh x \sin y$
$\tanh(x \pm \mathrm{i}y)$	$\dfrac{\sinh 2x}{\cosh 2x + \cos 2y}$	$\pm\dfrac{\sin 2y}{\cosh 2x + \cos 2y}$

7. Absolute Values and Arguments of the Trigonometric and Hyperbolic Functions (See Table 14.2)

Table 14.2 Absolute values and arguments of the trigonometric and hyperbolic functions

Function $w = f(x+iy)$	**Absolute value** $\lvert w \rvert$	**Argument arg** w
$\sin(x \pm \mathrm{i}y)$	$\sqrt{\sin^2 x + \sinh^2 y}$	$\pm\arctan(\cot x\ \tanh y)$
$\cos(x \pm \mathrm{i}y)$	$\sqrt{\cos^2 x + \sinh^2 y}$	$\mp\arctan(\tan x\ \tanh y)$
$\sinh(x \pm \mathrm{i}y)$	$\sqrt{\sinh^2 x + \sin^2 y}$	$\pm\arctan(\coth x\ \tan y)$
$\cosh(x \pm \mathrm{i}y)$	$\sqrt{\sinh^2 x + \cos^2 y}$	$\pm\arctan(\tanh x\ \tan y)$

14.5.3 Description of Curves in Complex Form

A complex function of one real variable t can be represented in parameter form:

$$z = x(t) + \mathrm{i}y(t) = f(t). \quad (14.91)$$

As t changes, the points z draw a curve $z(t)$. Now, the equations and the corresponding graphical representations of the line, circle, hyperbola, ellipse, and logarithmic spiral are represented.

1. Straight Line

a) Line through a point (z_1, φ) (φ is the angle with the x-axis, **Fig. 14.49a):**

$$z = z_1 + te^{\mathrm{i}\varphi}. \tag{14.92a}$$

b) Line through two points $z_1,\ z_2$ **(Fig. 14.49b):**

$$z = z_1 + t(z_2 - z_1). \tag{14.92b}$$

2. Circle

a) Circle with radius r, center at the point $z_0 = 0$ **(Fig. 14.50a):**

$$z = re^{\mathrm{i}t} \quad (|z| = r). \tag{14.93a}$$

b) Circle with radius r, center at the point z_0 **(Fig. 14.50b):**

$$z = z_0 + re^{\mathrm{i}t} \quad (|z - z_0| = r). \tag{14.93b}$$

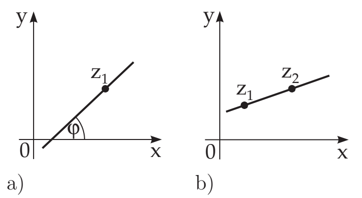


Figure 14.49

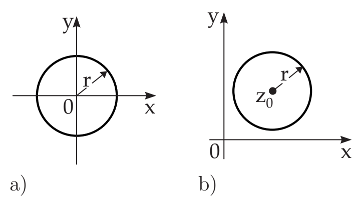


Figure 14.50

3. Ellipse

a) Ellipse, Normal Form $\dfrac{x^2}{a^2} + \dfrac{y^2}{b^2} = 1$ **(Fig. 14.51a):**

$$z = a\cos t + \mathrm{i}b\sin t \ (14.94\mathrm{a}) \ \text{ or } z = c\,e^{\mathrm{i}t} + d\,e^{-\mathrm{i}t} \ (14.94\mathrm{b}) \text{ with } c = \frac{a+b}{2},\quad d = \frac{a-b}{2}, \tag{14.94c}$$

i.e., c and d are arbitrary real numbers.

b) Ellipse, General Form (Fig. 14.51b): The center is at z_1, the axes are rotated by an angle.

$$z = z_1 + ce^{\mathrm{i}t} + de^{-\mathrm{i}t}. \tag{14.95}$$

Here c and d are arbitrary complex numbers, they determine the length of the axis of the ellipse and the angle of rotation.

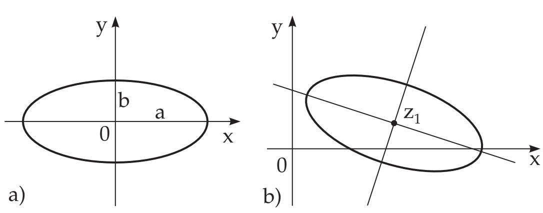


Figure 14.51

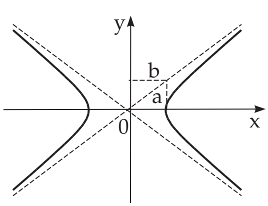


Figure 14.52

4. Hyperbola, Normal Form $\dfrac{x^2}{a^2} - \dfrac{y^2}{b^2} = 1$ (Fig. 14.52):

$$z = a\cosh t + \mathrm{i}b\sinh t \qquad (14.96a)$$

or

$$z = ce^t + \bar{c}e^{-t}, \qquad (14.97)$$

where c and $\bar{c}$ are conjugate complex numbers:

$$c = \frac{a + \mathrm{i}b}{2}, \quad \bar{c} = \frac{a - \mathrm{i}b}{2}. \qquad (14.98)$$

5. Logarithmic Spiral (Fig. 14.53):

$$z = a\, e^{\mathrm{i}bt}, \qquad (14.99)$$

where a and b are arbitrary complex numbers.

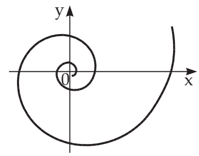


Figure 14.53

14.6 Elliptic Functions

14.6.1 Relation to Elliptic Integrals

Integrals in the form (8.22) with integrands $R(x, \sqrt{P(x)})$ can not be integrated in closed form if $P(x)$ is a polynomial of degree three or four, except in some special cases, but they are calculated numerically as elliptic integrals (see 8.1.4.3, p. 490). The inverse functions of elliptic integrals are the *elliptic functions*. They are similar to the trigonometric functions and they can be considered as their generalization. As an illustration the special case

$$\int\limits_0^u (1 - t^2)^{-\frac{1}{2}}\, dt = x \quad (|u| \le 1). \qquad (14.100)$$

is considered.

a) There is a relation between the trigonometric function $u = \sin x$ and the principal value of its inverse function

$$u = \sin x \Leftrightarrow x = \arcsin u \quad \text{for} \quad -\frac{\pi}{2} \le x \le \frac{\pi}{2}, \ -1 \le u \le 1, \qquad (14.101)$$

b) the integral (14.100) is equal to $\arcsin u$. The sine function can be considered as the inverse function of the integral (14.100). Analogies are valid for the elliptic integrals.

■ The period of a *mathematical pendulum*, with mass m, hanging on a non-elastic weightless thread of length l **(Fig. 14.54)**, can be calculated by a second-order non-linear differential equation. This equation follows from the balance of the forces acting on the mass of the pendulum:

$$\frac{d^2\vartheta}{dt^2} + \frac{g}{l}\sin\vartheta = 0 \text{ with } \vartheta(0) = \vartheta_0,\ \dot{\vartheta}(0) = 0 \text{ or } \frac{d}{dt}\left[\left(\frac{d\vartheta}{dt}\right)^2\right] = 2\,\frac{g}{l}\,\frac{d}{dt}(\cos\vartheta). \qquad (14.102a)$$

The relation between the length l and the amplitude s from the normal position is $s = l\vartheta$, so $\dot{s} = l\dot{\vartheta}$ and $\ddot{s} = l\ddot{\vartheta}$ hold. The force acting on the mass is $F = mg$, where g is the acceleration due to gravity (see **Table 21.2**, p. 1053), and it is decomposed into a normal component F_N and a tangential component F_T with respect to its path **(Fig. 14.54)**. The normal component $F_N = mg\cos\vartheta$ is balanced by the thread stress. Since it is perpendicular to the direction of motion, it has no effect to the equation of motion. The tangential component F_T yields the acceleration of the motion. $F_T = m\ddot{s} = ml\ddot{\vartheta} = -mg\sin\vartheta$. The tangential component always points in the direction of the normal position.
By separation of variables follows:

$$t - t_0 = \sqrt{\frac{l}{g}}\int_0^{\vartheta} \frac{d\Theta}{\sqrt{2(\cos\Theta - \cos\vartheta_0)}}. \qquad (14.102b)$$

Here, t_0 denotes the time for which the pendulum is in the deepest position for the first time, i.e., where $\vartheta(t_0) = 0$ holds. Θ denotes the integration variable. After some transformations and with the substitutions $\sin\frac{\Theta}{2} = k\sin\psi\,,\quad k = \sin\frac{\vartheta_0}{2}$ follows

$$t - t_0 = \sqrt{\frac{l}{g}}\int_0^{\varphi}\frac{d\psi}{\sqrt{1-k^2\sin^2\psi}} = \sqrt{\frac{l}{g}}\,F(k,\varphi)\,. \tag{14.102c}$$

Here $F(k,\varphi)$ is an elliptic integral of the first kind (see (8.25a), p. 490). The angle of deflection $\vartheta = \vartheta(t)$ is a periodic function of period $2T$ with

$$T = \sqrt{\frac{l}{g}}\,F\left(k,\frac{\pi}{2}\right) = \sqrt{\frac{l}{g}}K\,, \tag{14.102d}$$

where K represents a complete elliptic integral of the first kind (**Table 21.9**, p. 1103). T denotes the *period* of the pendulum, i.e., the time between two consecutive extreme positions for which $\frac{d\vartheta}{dt} = 0$. If the amplitude is small, i.e., $\sin\vartheta \approx \vartheta$, then $T = 2\pi\sqrt{l/g}$ holds.

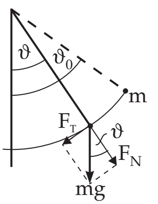


Figure 14.54

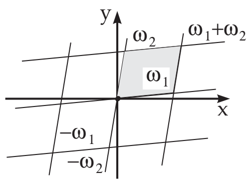


Figure 14.55

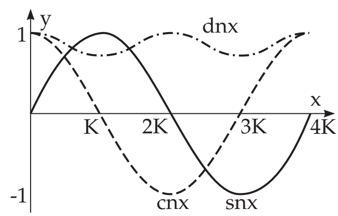


Figure 14.56

14.6.2 Jacobian Functions

1. Definition

It follows for $0 < k < 1$ from the representation (8.24a) and (8.25a) (see 8.1.4.3, p. 490) for the elliptic integral of the first kind $F(k,\varphi)$ that

$$\frac{dF}{d\varphi} = (1 - k^2\sin^2\varphi)^{-\frac{1}{2}} > 0, \tag{14.103}$$

i.e., $F(k,\varphi)$ is strictly monotone with respect to φ, so the inverse function

$$\varphi = \mathrm{am}(k,u) = \varphi(u) \qquad (14.104\text{a}) \qquad\qquad \text{of}\quad u = \int\limits_0^{\varphi}\frac{d\psi}{\sqrt{1-k^2\sin^2\psi}} = u(\varphi) \qquad (14.104\text{b})$$

exists. It is called the *amplitude function*. The so-called *Jacobian functions* are defined as:

$$\mathrm{sn}u = \sin\varphi = \sin\mathrm{am}(k,u) \quad \text{(amplitude sine)}, \tag{14.105a}$$

$$\mathrm{cn}u = \cos\varphi = \cos\mathrm{am}(k,u) \quad \text{(amplitude cosine)}, \tag{14.105b}$$

$$\mathrm{dn}u = \sqrt{1-k^2\mathrm{sn}^2u} \qquad \text{(amplitude delta)}. \tag{14.105c}$$

2. Meromorphic and Double Periodic Functions

The Jacobian functions can be continued analytically in the z plane. The functions $\mathrm{sn}z$, $\mathrm{cn}z$, and $\mathrm{dn}z$ are then *meromorphic* functions (see 14.3.5.2, p. 753), i.e., they have only poles as singularities. Besides,

they are *double periodic*: Each of these functions $f(z)$ has exactly two periods ω_1 and ω_2 with

$$f(z+\omega_1)=f(z), \quad f(z+\omega_2)=f(z). \tag{14.106}$$

Here, ω_1 and ω_2 are two arbitrary complex numbers, whose ratio is not real. The general formula

$$f(z+m\omega_1+n\omega_2)=f(z) \tag{14.107}$$

follows from (14.106), and here m and n are arbitrary integers. Meromorphic double periodic functions are called *elliptic functions*. The set

$$\{z_0+\alpha_1\omega_1+\alpha_2\omega_2\colon\ 0\le\alpha_1,\alpha_2<1\}, \tag{14.108}$$

with an arbitrary fixed $z_0\in\mathbb{C}$, is called the *period parallelogram* of the elliptic function. If this function is bounded in the whole period parallelogram **(Fig. 14.55)**, then it is a constant.

■ The Jacobian functions (14.105a) and (14.105b) are elliptic functions. The amplitude function (14.104a) is not an elliptic function.

3. Properties of the Jacobian functions

The properties of the Jacobian functions given in **Table 14.3** can be got by the substitutions

$$k'^2=1-k^2, \quad K'=F\left(k',\frac{\pi}{2}\right), \ K=F\left(k,\frac{\pi}{2}\right), \tag{14.109}$$

where m and n are arbitrary integers.

Table 14.3 Periods, roots and poles of Jacobian functions

	Periods ω_1,ω_2	**Roots**	**Poles**
sn$z$	$4K, 2\mathrm{i}K'$	$2mK+2n\mathrm{i}K'$	$2mK+(2n+1)\mathrm{i}K'$ (for sn$z$, cn$z$, dn$z$)
cn$z$	$4K, 2(K+\mathrm{i}K')$	$(2m+1)K+2n\mathrm{i}K'$	
dn$z$	$2K, 4\mathrm{i}K'$	$(2m+1)K+(2n+1)\mathrm{i}K'$	

The shape of sn$z$, cn$z$, and dn$z$ can be found in **Fig. 14.56**. The following relations are valid for the Jacobian functions except at the poles:

1. $$\mathrm{sn}^2z+\mathrm{cn}^2z=1, \quad k^2\mathrm{sn}^2z+\mathrm{dn}^2z=1, \tag{14.110}$$

2. $$\mathrm{sn}(u+v)=\frac{(\mathrm{sn}u)(\mathrm{cn}v)(\mathrm{dn}v)+(\mathrm{sn}v)(\mathrm{cn}u)(\mathrm{dn}u)}{1-k^2(\mathrm{sn}^2u)(\mathrm{sn}^2v)}, \tag{14.111a}$$

$$\mathrm{cn}(u+v)=\frac{(\mathrm{cn}u)(\mathrm{cn}v)-(\mathrm{sn}u)(\mathrm{dn}u)(\mathrm{sn}v)(\mathrm{dn}v)}{1-k^2(\mathrm{sn}^2u)(\mathrm{sn}^2v)}, \tag{14.111b}$$

$$\mathrm{dn}(u+v)=\frac{(\mathrm{dn}u)(\mathrm{dn}v)-k^2(\mathrm{sn}u)(\mathrm{cn}u)(\mathrm{sn}v)(\mathrm{cn}v)}{1-k^2(\mathrm{sn}^2u)(\mathrm{sn}^2v)}, \tag{14.111c}$$

3. $$(\mathrm{sn}z)'=(\mathrm{cn}z)(\mathrm{dn}z), \tag{14.112a}$$ $$(\mathrm{cn}z)'=-(\mathrm{sn}z)(\mathrm{dn}z), \tag{14.112b}$$

$$(\mathrm{dn}z)'=-k^2(\mathrm{sn}z)(\mathrm{cn}z). \tag{14.112c}$$

For further properties of the Jacobian functions and further elliptic functions see [14.8], [14.12].

14.6.3 Theta Functions

The *theta functions* can be used to evaluate the Jacobian functions:

$$\vartheta_1(z,q)=2q^{\frac14}\sum_{n=0}^{\infty}(-1)^nq^{n(n+1)}\sin(2n+1)z, \tag{14.113a}$$

$$\vartheta_2(z,q) = 2q^{\frac{1}{4}} \sum_{n=0}^{\infty} q^{n(n+1)} \cos(2n+1)z, \tag{14.113b}$$

$$\vartheta_3(z,q) = 1 + 2\sum_{n=1}^{\infty} q^{n^2} \cos 2nz, \tag{14.113c}$$

$$\vartheta_4(z,q) = 1 + 2\sum_{n=1}^{\infty} (-1)^n q^{n^2} \cos 2nz. \tag{14.113d}$$

If $|q| < 1$ (q complex) holds, then the series (14.113a)–(14.113d) are convergent for every complex argument z. In the case of a constant q one uses the following brief notations:

$$\vartheta_k(z) := \vartheta_k(\pi z, q) \quad (k = 1,2,3,4). \tag{14.114}$$

Then, the Jacobian functions have the representations:

$$\mathrm{sn}z = 2K\frac{\vartheta_4(0)}{\vartheta_1'(0)}\frac{\vartheta_1\left(\dfrac{z}{2K}\right)}{\vartheta_4\left(\dfrac{z}{2K}\right)}, \quad (14.115a) \qquad \mathrm{cn}z = \frac{\vartheta_4(0)}{\vartheta_2(0)}\frac{\vartheta_2\left(\dfrac{z}{2K}\right)}{\vartheta_4\left(\dfrac{z}{2K}\right)}, \tag{14.115b}$$

$$\mathrm{dn}z = \frac{\vartheta_4(0)}{\vartheta_3(0)}\frac{\vartheta_3\left(\dfrac{z}{2K}\right)}{\vartheta_4\left(\dfrac{z}{2K}\right)}, \quad (14.115c) \quad \text{with} \quad q = \exp\left(-\pi\frac{K'}{K}\right), \quad k = \left(\frac{\vartheta_2(0)}{\vartheta_3(0)}\right)^2 \tag{14.115d}$$

and K, K' are as in (14.109).

14.6.4 Weierstrass Functions

The functions

$$\wp(z) = \wp(z,\omega_1,\omega_2), \quad (14.116a) \qquad \zeta(z) = \zeta(z,\omega_1,\omega_2), \tag{14.116b}$$

$$\sigma(z) = \sigma(z,\omega_1,\omega_2), \tag{14.116c}$$

were introduced by Weierstrass, and here ω_1 and ω_2 represent two arbitrary complex numbers whose quotient is not real. One substitutes

$$\omega_{mn} = 2(m\omega_1 + n\omega_2), \tag{14.117a}$$

where m and n are arbitrary real numbers, and defines

$$\wp(z,\omega_1,\omega_2) = z^{-2} + {\sum_{m,n}}' \left[(z-\omega_{mn})^{-2} - \omega_{mn}^{-2}\right]. \tag{14.117b}$$

The accent behind the sum sign denotes that the value pair $m = n = 0$ is omitted. The function $\wp(z,\omega_1,\omega_2)$ has the following properties:

1. It is an elliptic function with periods ω_1 and ω_2.

2. The series (14.117b) is convergent for every $z \neq \omega_{mn}$.

3. The function $\wp(z,\omega_1,\omega_2)$ satisfies the differential equation

$$\wp'^2 = 4\wp^3 - g_2\wp - g_3 \quad (14.118a) \qquad \text{with } g_2 = 60{\sum_{m,n}}'\omega_{mn}^{-4}, \quad g_3 = 140{\sum_{m,n}}'\omega_{mn}^{-6}. \tag{14.118b}$$

The quantities g_2 and g_3 are called the *invariants* of $\wp(z,\omega_1,\omega_2)$.

4. The function $u = \wp(z, \omega_1, \omega_2)$ is the inverse function of the integral

$$z = \int\limits_u^\infty \frac{dt}{\sqrt{4t^3 - g_2 t - g_3}}\,. \tag{14.119}$$

5.
$$\wp(u+v) = \frac{1}{4}\left[\frac{\wp'(u) - \wp'(v)}{\wp(u) - \wp(v)}\right]^2 - \wp(u) - \wp(v). \tag{14.120}$$

The Weierstrass functions

$$\zeta(z) = z^{-1} + \sum_{m,n}{}' \left[(z - \omega_{mn})^{-1} + \omega_{mn}^{-1} + \omega_{mn}^{-2} z\right], \tag{14.121a}$$

$$\sigma(z) = z\exp\left(\int_0^z \left[\zeta(t) - t^{-1}\right] dt\right) = z \prod_{m,n}{}' \left(1 - \frac{z}{\omega_{mn}}\right) \exp\left(\frac{z}{\omega_{mn}} + \frac{z^2}{2\omega_{mn}^2}\right) \tag{14.121b}$$

are not double periodic, so they are not elliptic functions. The following relations are valid:

1. $$\zeta'(z) = -\wp(z), \;\; \zeta(z) = (\ln \sigma(z))\,, \tag{14.122}$$

2. $$\zeta(-z) = -\zeta(z), \;\; \sigma(-z) = -\sigma(z), \tag{14.123}$$

3. $$\zeta(z + 2\omega_1) = \zeta(z) + 2\zeta(\omega_1), \quad \zeta(z + 2\omega_2) = \zeta(z) + 2\zeta(\omega_2), \tag{14.124}$$

4. $$\zeta(u+v) = \zeta(u) + \zeta(v) + \frac{1}{2}\frac{\wp'(u) - \wp'(v)}{\wp(u) - \wp(v)}\,. \tag{14.125}$$

5. Every elliptic function is a rational function of the Weierstrass functions $\wp(z)$ and $\zeta(z)$.

15 Integral Transformations

15.1 Notion of Integral Transformation

15.1.1 General Definition of Integral Transformations

AnIntegral transformation* is a correspondence between two functions $f(t)$ and $F(p)$ in the form

$$F(p) = \int_{-\infty}^{+\infty} K(p,t) f(t)\, dt. \tag{15.1a}$$

The function $f(t)$ is called the *original function*, its domain is the *original space*. The function $F(p)$ is called the *transform*, its domain is the *image space*.
The function $K(p,t)$ is called the *kernel* of the transformation. In general, t is a real variable, and $p = \sigma + \mathrm{i}\omega$ is a complex variable.
The shorter notation can be used by introducing the symbol $\mathcal{T}$ for the integral transformation with kernel $K(p,t)$:

$$F(p) = \mathcal{T}\{f(t)\}. \tag{15.1b}$$

Then, (15.1b) is called $\mathcal{T}$ transformation.

15.1.2 Special Integral Transformations

Different kernels $K(p,t)$ and different original spaces yield different integral transformations. The most widely known transformations are the Laplace transformation, the Laplace-Carson transformation, and the Fourier transformation. In this book an overview is given about the integral transformations of functions of one variable (see also **Table 15.1**). More recently, some additional transformations have been introduced for use in pattern recognition and in characterizing signals, such as the Wavelet transformation, the Gabor transformation and the Walsh transformation (see 15.6, p. 800ff.).

15.1.3 Inverse Transformations

The inverse transformation of a transform into the original function has special importance in applications. With the symbol $\mathcal{T}^{-1}$ the inverse integral transformation of (15.1a) is

$$f(t) = \mathcal{T}^{-1}\{F(p)\}. \tag{15.2a}$$

The operator $\mathcal{T}^{-1}$ is called the *inverse operator* of $\mathcal{T}$, so

$$\mathcal{T}^{-1}\{\mathcal{T}\{f(t)\}\} = f(t). \tag{15.2b}$$

The determination of the inverse transformation means the solution of the integral equation (15.1a), where the function $F(p)$ is given and function $f(t)$ is to be determined. If there is a solution, then it can be written in the form

$$f(t) = \mathcal{T}^{-1}\{F(p)\}. \tag{15.2c}$$

The explicit determination of *inverse operators* for different integral transformations, i.e., for different kernels $K(p,t)$, belongs to the fundamental problems of the theory of integral transformations. The user can solve practical problems by using the given correspondences between transforms and original functions in the corresponding tables (**Table 21.13**, p. 1109, **Table 21.14**, p. 1114, and **Table 21.15**, p. 1128).

15.1.4 Linearity of Integral Transformations

If $f_1(t)$ and $f_2(t)$ are transformable functions, then

$$\mathcal{T}\{k_1 f_1(t) + k_2 f_2(t)\} = k_1 \mathcal{T}\{f_1(t)\} + k_2 \mathcal{T}\{f_2(t)\}, \tag{15.3}$$

where k_1 and k_2 are arbitrary numbers. That is, an integral transformation represents a linear operation on the set T of the $\mathcal{T}$-transformable functions.

I.N. Bronshtein et al., *Handbook of Mathematics*,
DOI 10.1007/978-3-662-46221-8_15

Table 15.1 Overview of integral transformations of functions of one variable

Transformation	**Kernel $K(p,t)$**	**Symbol**	**Remark**
Laplace transformation	$\begin{cases} 0 & \text{for } t<0 \\ e^{-pt} & \text{for } t>0 \end{cases}$	$\mathcal{L}\{f(t)\} = \int_0^{\infty} e^{-pt} f(t)dt$	$p=\sigma+\mathrm{i}\omega$
Two-sided Laplace transformation	e^{-pt}	$\mathcal{L}_{\mathrm{II}}\{f(t)\} = \int_{-\infty}^{+\infty} e^{-pt} f(t)dt$	$\mathcal{L}_{\mathrm{II}}\{f(t)\mathit{1}(t)\} = \mathcal{L}\{f(t)\}$ where $\mathit{1}(t) = \begin{cases} 0 \text{ for } t<0 \\ 1 \text{ for } t>0 \end{cases}$
Finite Laplace transformation	$\begin{cases} 0 & \text{for } t<0 \\ e^{-pt} & \text{for } 0<t<a \\ 0 & \text{for } t>a \end{cases}$	$\mathcal{L}_a\{f(t)\} = \int_0^{a} e^{-pt} f(t)dt$	
Laplace-Carson transformation	$\begin{cases} 0 & \text{for } t<0 \\ pe^{-pt} & \text{for } t>0 \end{cases}$	$\mathcal{C}\{f(t)\} = \int_0^{\infty} pe^{-pt} f(t)dt$	The Carson transformation can also be a two-sided and finite transformation.
Fourier transformation	$e^{-\mathrm{i}\omega t}$	$\mathcal{F}\{f(t)\} = \int_{-\infty}^{+\infty} e^{-\mathrm{i}\omega t} f(t)dt$	$p=\sigma+\mathrm{i}\omega \quad \sigma=0$
One-sided Fourier transformation	$\begin{cases} 0 & \text{for } t<0 \\ e^{-\mathrm{i}\omega t} & \text{for } t>0 \end{cases}$	$\mathcal{F}_{\mathrm{I}}\{f(t)\} = \int_0^{\infty} e^{-\mathrm{i}\omega t} f(t)dt$	$p=\sigma+\mathrm{i}\omega \quad \sigma=0$
Finite Fourier transformation	$\begin{cases} 0 & \text{for } t<0 \\ e^{-\mathrm{i}\omega t} & \text{for } 0<t<a \\ 0 & \text{for } t>a \end{cases}$	$\mathcal{F}_a\{f(t)\} = \int_0^{a} e^{-\mathrm{i}\omega t} f(t)dt$	$p=\sigma+\mathrm{i}\omega \quad \sigma=0$
Fourier cosine transformation	$\begin{cases} 0 & \text{for } t<0 \\ \mathrm{Re}\,[e^{\mathrm{i}\omega t}] & \text{for } t>0 \end{cases}$	$\mathcal{F}_c\{f(t)\} = \int_0^{\infty} f(t)\cos\omega t\,dt$	$p=\sigma+\mathrm{i}\omega \quad \sigma=0$
Fourier sine transformation	$\begin{cases} 0 & \text{for } t<0 \\ \mathrm{Im}\,[e^{\mathrm{i}\omega t}] & \text{for } t>0 \end{cases}$	$\mathcal{F}_s\{f(t)\} = \int_0^{\infty} f(t)\sin\omega t\,dt$	$p=\sigma+\mathrm{i}\omega \quad \sigma=0$
Mellin transformation	$\begin{cases} 0 & \text{for } t<0 \\ t^{p-1} & \text{for } t>0 \end{cases}$	$\mathcal{M}\{f(t)\} = \int_0^{\infty} t^{p-1} f(t)dt$	
Hankel transformation of order ν	$\begin{cases} 0 & \text{for } t<0 \\ tJ_\nu(\sigma t) & \text{for } t>0 \end{cases}$	$\mathcal{H}_\nu\{f(t)\} = \int_0^{\infty} tJ_\nu(\sigma t) f(t)dt$	$p=\sigma+\mathrm{i}\omega \quad \omega=0$ $J_\nu(\sigma t)$ is the ν-th order Bessel function of the first kind.
Stieltjes transformation	$\begin{cases} 0 & \text{for } t<0 \\ \dfrac{1}{p+t} & \text{for } t>0 \end{cases}$	$\mathcal{S}\{f(t)\} = \int_0^{\infty} \dfrac{f(t)}{p+t} dt$	

15.1.5 Integral Transformations for Functions of Several Variables

Integral transformations for functions of several variables are also called *multiple integral transformations* (see [15.13]). The best-known ones are the double Laplace transformation, i.e., the Laplace transformation for functions of two variables, the double Laplace-Carson transformation and the double Fourier transformation. The definition of the double Laplace transformation is

$$F(p,q) = \mathcal{L}^2\{f(x,y)\} \equiv \int\limits_{x=0}^{\infty}\int\limits_{y=0}^{\infty} e^{-px-qy} f(x,y)\,dx\,dy. \tag{15.4}$$

The symbol $\mathcal{L}$ denotes the Laplace transformation for functions of one variable (see **Table 15.1**).

15.1.6 Applications of Integral Transformations

1. Fields of Applications

Besides the great theoretical importance that integral transformations have in such basic fields of mathematics as the theory of integral equations and the theory of linear operators, they have a large field of applications in the solutions of practical problems in physics and engineering. Methods with applications of integral transformations are often called *operator methods*. They are suitable to solve ordinary and partial differential equations, integral equations and difference equations.

2. Scheme of the Operator Method

The general scheme to the use of an operator method with an integral transformation is represented in **Fig. 15.1**. One gets the solution of a problem not directly from the original defining equation; first an integral transformation is applied. The inverse transformation of the solution of the transformed equation gives the solution of the original problem.

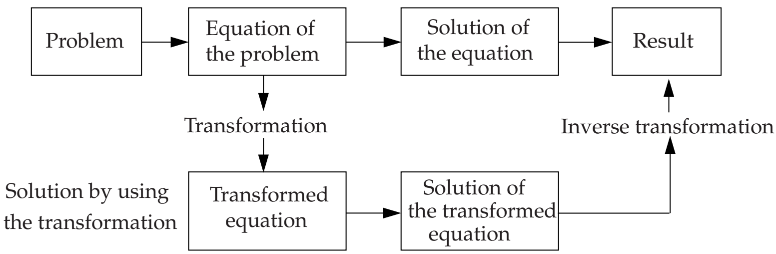


Figure 15.1

The application of the operator method to solve ordinary differential equations consists of the following three steps:

1. Transition from a differential equation of an unknown function to an equation of its transform.

2. Solution of the transformed equation in the image space. The transformed equation is usually no longer a differential equation, but an algebraic equation.

3. Inverse transformation of the transform with help of $\mathcal{T}^{-1}$ into the original space, i.e., determination of the solution of the original problem.

The major difficulty of the operator method is usually not the solution of the transformed equation, but the transformation of the function and the inverse transformation.

15.2 Laplace Transformation

15.2.1 Properties of the Laplace Transformation

15.2.1.1 Laplace Transformation, Original and Image Space

1. Definition of the Laplace Transformation

The Laplace transformation

$$\mathcal{L}\{f(t)\} = \int_0^\infty e^{-pt} f(t)\, dt = F(p) \tag{15.5}$$

assigns a function $F(p)$ of a complex variable p to a function $f(t)$ of a real variable t, if the given improper integral exists. $f(t)$ is called the *original function*, $F(p)$ is called the *transform* of $f(t)$. In the further discussion it is assumed that the improper integral exists if the original function $f(t)$ is piecewise smooth in its domain $t \geq 0$, in the so called *original space*, and for $t \to \infty$, $|f(t)| \leq Ke^{\alpha t}$ holds with certain constants $K > 0$, $\alpha > 0$. The domain of the transform $F(p)$ is called the *image space*.
In the literature the Laplace transformation is often found also in the Wagner or Laplace-Carson form

$$\mathcal{L}_W\{f(t)\} = p \int_0^\infty e^{-pt} f(t)\, dt = p\, F(p). \tag{15.6}$$

2. Convergence

The Laplace integral $\mathcal{L}\{f(t)\}$ converges in the half-plane $\operatorname{Re} p > \alpha$ **(Fig. 15.2)**. The transform $F(p)$ is an analytic function with the properties:

1. $$\lim_{\operatorname{Re} p \to \infty} F(p) = 0. \tag{15.7a}$$

This property is a necessary condition for $F(p)$ to be a transform.

2. $$\lim_{\substack{p \to 0 \\ (p \to \infty)}} p\, F(p) = A, \tag{15.7b}$$

if the original function $f(t)$ has a finite limit $\lim\limits_{\substack{t \to \infty \\ (t \to 0)}} f(t) = A$.

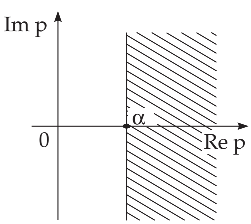


Figure 15.2

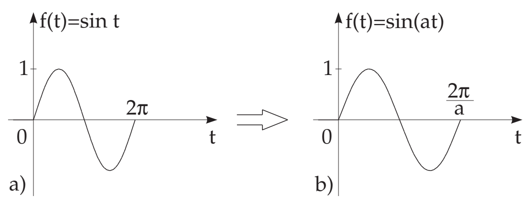


Figure 15.3

3. Inverse Laplace Transformation

One can retrieve the original function from the transform with the formula

$$\mathcal{L}^{-1}\{F(p)\} = \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} e^{pt} F(p)\, dp = \begin{cases} f(t) & \text{for } t > 0, \\ 0 & \text{for } t < 0. \end{cases} \tag{15.8}$$

The path of integration of this complex integral is a line $\mathrm{Re}\, p = c$ parallel to the imaginary axis, where $\mathrm{Re}\, p = c > \alpha$. If the function $f(t)$ has a jump at $t = 0$, i.e., $\lim_{t \to +0} f(t) \neq 0$, then the integral has the value $\frac{1}{2} f(+0)$ there.

15.2.1.2 Rules for the Evaluation of the Laplace Transformation

The rules for evaluation are the mappings of operations in the original domain into operations in the transform space.
Hereafter the original functions will be denoted by lowercase letters, the transforms are denoted by the corresponding capital letters.

1. Addition or Linearity Law

The Laplace transform of a linear combination of functions is the same linear combination of the Laplace transforms, if they exist. With constants $\lambda_1, \ldots, \lambda_n$ that is

$$\mathcal{L}\{\lambda_1 f_1(t) + \lambda_2 f_2(t) + \cdots + \lambda_n f_n(t)\} = \lambda_1 F_1(p) + \lambda_2 F_2(p) + \cdots + \lambda_n F_n(p). \qquad (15.9)$$

2. Similarity Laws

The Laplace transform of $f(at)$ $(a > 0, a$ real) is the Laplace transform of the original function divided by a and with the argument p/a:

$$\mathcal{L}\{f(at)\} = \frac{1}{a} F\left(\frac{p}{a}\right) \quad (a > 0,\ \text{real}). \qquad (15.10a)$$

Analogously for the inverse transformation

$$\mathcal{L}^{-1}\{F(ap)\} = \frac{1}{a} f\left(\frac{t}{a}\right) \quad (a > 0). \qquad (15.10b)$$

Fig. 15.3 shows the application of the similarity laws for a sine function.

■ Determination of the Laplace transform of $f(t) = \sin(\omega t)$. The transform of the sine function is $\mathcal{L}\{\sin(t)\} = F(p) = 1/(p^2 + 1)$. Application of the similarity law gives $\mathcal{L}\{\sin(\omega t)\} = \frac{1}{\omega} F(p/\omega) = \frac{1}{\omega} \frac{1}{(p/\omega)^2 + 1} = \frac{\omega}{p^2 + \omega^2}$.

3. Translation Laws

1. Shifting to the Right The Laplace transform of an original function shifted to the right by a $(a > 0)$ is equal to the Laplace transform of the non-shifted original function multiplied by the factor e^{-ap}:

$$\mathcal{L}\{f(t - a)\} = e^{-ap} F(p). \qquad (15.11a)$$

2. Shifting to the Left The Laplace transform of an original function shifted to the left by a is equal to e^{ap} multiplied by the difference of the transform of the non-shifted function and the integral $\int_0^a f(t)\, e^{-pt}\, dt$:

$$\mathcal{L}\{f(t + a)\} = e^{ap} \left[F(p) - \int_0^a e^{-pt} f(t)\, dt \right]. \qquad (15.11b)$$

Figs. 15.4 and **15.5** show the cosine function shifted to the right and a line shifted to the left.

4. Frequency Shift Theorem

The Laplace transform of an original function multiplied by e^{-bt} is equal to the Laplace transform with the argument $p + b$ (b is an arbitrary complex number):

$$\mathcal{L}\{e^{-bt} f(t)\} = F(p + b). \qquad (15.12)$$

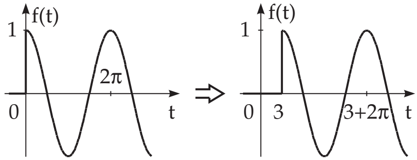


Figure 15.4

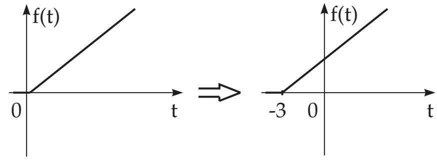


Figure 15.5

5. Differentiation in the Original Space

If the derivatives $f'(t), f''(t), \ldots, f^{(n)}(t)$ exist for $t > 0$ and the highest derivative of $f(t)$ has a transform, then the lower derivatives of $f(t)$ and also $f(t)$ have a transform, and:

$$\left.\begin{array}{l} \mathcal{L}\{f'(t)\} = p\,F(p) - f(+0), \\ \mathcal{L}\{f''(t)\} = p^2\,F(p) - f(+0)\,p - f'(+0), \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \mathcal{L}\{f^{(n)}(t)\} = p^n\,F(p) - f(+0)\,p^{n-1} - f'(+0)\,p^{n-2} - \cdots \\ \qquad - f^{(n-2)}(+0)\,p - f^{(n-1)}(+0) \text{ with} \\ f^{(\nu)}(+0) = \lim\limits_{t\to+0} f^{(\nu)}(t). \end{array}\right\} \tag{15.13}$$

Equation (15.13) gives the following representation of the Laplace integral, which can be used for approximating the Laplace integral:

$$\mathcal{L}\{f(t)\} = \frac{f(+0)}{p} + \frac{f'(+0)}{p^2} + \frac{f''(+0)}{p^3} + \cdots + \frac{f^{(n-1)}(+0)}{p^{n-1}} + \frac{1}{p^n}\mathcal{L}\{f^{(n)}(t)\}. \tag{15.14}$$

6. Differentiation in the Image Space

$$\mathcal{L}\{t^n f(t)\} = (-1)^n F^{(n)}(p). \tag{15.15}$$

The n-th derivative of the transform is equal to the Laplace transform of the $(-t)^n$-th multiple of the original function $f(t)$:

$$\mathcal{L}\{(-1)^n\,t^n\,f(t)\} = F^{(n)}(p) \qquad (n = 1, 2, \ldots)\,. \tag{15.16}$$

7. Integration in the Original Space

The transform of an integral of the original function is equal to $1/p^n$ $(n > 0)$ multiplied by the transform of the original function:

$$\mathcal{L}\left\{\int\limits_0^t d\tau_1 \int\limits_0^{\tau_1} d\tau_2 \ldots \int\limits_0^{\tau_{n-1}} f(\tau_n)\,d\tau_n\right\} = \frac{1}{(n-1)\,!}\,\mathcal{L}\left\{\int\limits_0^t (t-\tau)^{(n-1)}\,f(\tau)\,d\tau\right\} = \frac{1}{p^n}F(p). \tag{15.17a}$$

In the special case of the ordinary simple integral

$$\mathcal{L}\left\{\int\limits_0^t f(\tau)\,d\tau\right\} = \frac{1}{p}F(p) \tag{15.17b}$$

holds. In the original space, differentiation and integration act in converse ways if the initial values are zeros.

8. Integration in the Image Space

$$\mathcal{L}\left\{\frac{f(t)}{t^n}\right\} = \int\limits_p^\infty dp_1 \int\limits_{p_1}^\infty dp_2 \ldots \int\limits_{p_{n-1}}^\infty F(p_n)\,dp_n = \frac{1}{(n-1)\,!}\int\limits_p^\infty (z-p)^{n-1} F(z)\,dz. \tag{15.18}$$

This formula is valid only if $f(t)/t^n$ has a Laplace transform. For this purpose, $f(x)$ must tend to zero fast enough as $t \to 0$. The path of integration can be any ray starting at p, which forms an acute angle with the positive half of the real axis.

9. Division Law

In the special case of $n = 1$ of (15.18)

$$\mathcal{L}\left\{\frac{f(t)}{t}\right\} = \int_p^{\infty} F(z)\,dz \tag{15.19}$$

holds. For the existence of the integral (15.19), the limit $\lim\limits_{t\to 0} \dfrac{f(t)}{t}$ must also exist.

10. Differentiation and Integration with Respect to a Parameter

$$\mathcal{L}\left\{\frac{\partial f(t,\alpha)}{\partial \alpha}\right\} = \frac{\partial F(p,\alpha)}{\partial \alpha}, \quad (15.20a) \qquad \mathcal{L}\left\{\int_{\alpha_1}^{\alpha_2} f(t,\alpha)\,d\alpha\right\} = \int_{\alpha_1}^{\alpha_2} F(p,\alpha)\,d\alpha. \quad (15.20b)$$

Sometimes one can calculate Laplace integrals from known integrals with the help of these formulas.

11. Convolution

1. Convolution in the Original Space The convolution of two functions $f_1(t)$ and $f_2(t)$ is the integral

$$f_1 * f_2 = \int_0^t f_1(\tau)\cdot f_2(t-\tau)\,d\tau\,. \tag{15.21}$$

Equation (15.21) is also called the *one-sided convolution* in the interval $(0, t)$. A *two-sided convolution* occurs for the Fourier transformation (convolution in the interval $(-\infty, \infty)$ see 15.3.1.3, **9.**, p. 789). The convolution (15.21) has the properties

a) Commutative law: $f_1 * f_2 = f_2 * f_1$. (15.22a)

b) Associative law: $(f_1 * f_2) * f_3 = f_1 * (f_2 * f_3)$. (15.22b)

c) Distributive law: $(f_1 + f_2) * f_3 = f_1 * f_3 + f_2 * f_3$. (15.22c)

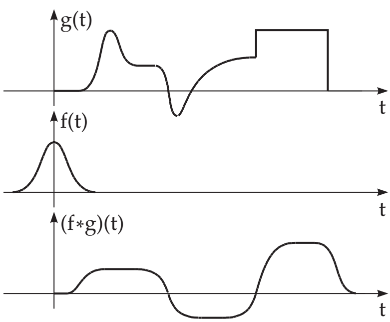


Figure 15.6

In the image domain, the usual multiplication corresponds to the convolution:

$$\mathcal{L}\{f_1 * f_2\} = F_1(p)\cdot F_2(p). \tag{15.23}$$

The convolution of two functions is shown in **Fig. 15.6**. One can apply the convolution theorem to determine the original function:

a) Factoring the transform $F(p) = F_1(p)\cdot F_2(p)$.

b) Determining the original functions $f_1(t)$ and $f_2(t)$ of the transforms $F_1(p)$ and $F_2(p)$ (from a table).

c) Determining the original function associated to $F(p)$ by convolution of $f_1(t)$ and $f_2(t)$ in the original space ($f(t) = f_1(t) * f_2(t)$).

2. Convolution in the Image Space (Complex Convolution)

$$\mathcal{L}\{f_1(t) \cdot f_2(t)\} = \begin{cases} \dfrac{1}{2\pi \mathrm{i}} \displaystyle\int_{x_1-\mathrm{i}\infty}^{x_1+\mathrm{i}\infty} F_1(z) \cdot F_2(p-z)\, dz, \\ \dfrac{1}{2\pi \mathrm{i}} \displaystyle\int_{x_2-\mathrm{i}\infty}^{x_2+\mathrm{i}\infty} F_1(p-z) \cdot F_2(z)\, dz. \end{cases} \tag{15.24}$$

The integration is performed along a line parallel to the imaginary axis. In the first integral, x_1 and p must be chosen so that z is in the half plane of convergence of $\mathcal{L}\{f_1\}$ and $p - z$ is in the half plane of convergence of $\mathcal{L}\{f_2\}$. The corresponding requirements must be valid for the second integral.

15.2.1.3 Transforms of Special Functions

1. Step Function

The unit jump at $t = t_0$ is called a step function **(Fig. 15.7)** (see also 14.4.3.2, **3.**, p. 757); it is also called the *Heaviside unit step function*:

$$u(t-t_0) = \begin{cases} 1 \text{ for } t > t_0, \\ 0 \text{ for } t < t_0 \end{cases} \quad (t_0 > 0). \tag{15.25}$$

■ **A:** $f(t) = u(t-t_0)\sin\omega t$, $\quad F(p) = e^{-t_0 p}\dfrac{\omega\,\cos\omega\, t_0 + p\,\sin\omega\, t_0}{p^2+\omega^2}$ **(Fig. 15.8)**.

■ **B:** $f(t) = u(t-t_0)\sin\omega\,(t-t_0)$, $\quad F(p) = e^{-t_0 p}\dfrac{\omega}{p^2+\omega^2}$ **(Fig. 15.9)**.

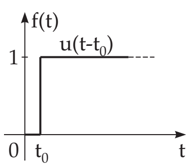


Figure 15.7

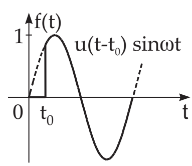


Figure 15.8

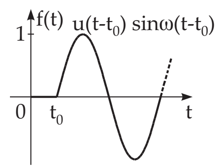


Figure 15.9

2. Rectangular Impulse

A rectangular impulse of height 1 and width T **(Fig. 15.10)** is composed by the superposition of two step functions in the form

$$u_T(t-t_0) = u(t-t_0) - u(t-t_0-T) = \begin{cases} 0 & \text{for } t < t_0, \\ 1 & \text{for } t_0 < t < t_0 + T, \\ 0 & \text{for } t > t_0 + T; \end{cases} \tag{15.26}$$

$$\mathcal{L}\{u_T(t-t_0)\} = \frac{e^{-t_0 p}(1-e^{-Tp})}{p}. \tag{15.27}$$

3. Impulse Function (Dirac δ Function)

(See also 12.9.5.4, p. 700.) The impulse function $\delta(t-t_0)$ can obviously be interpreted as a limit of the rectangular impulse of width T and height $1/T$ at the point $t = t_0$ **(Fig. 15.11)**:

$$\delta(t-t_0) = \lim_{T\to 0}\frac{1}{T}[\,u(t-t_0) - u(t-t_0-T)\,]. \tag{15.28}$$

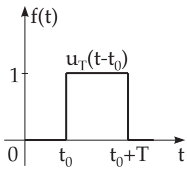


Figure 15.10

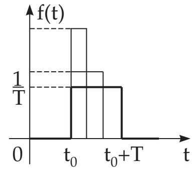


Figure 15.11

For a continuous function $h(t)$,

$$\int_a^b h(t)\,\delta(t-t_0)\,dt = \begin{cases} h(t_0), & \text{if } t_0 \text{ is inside } (a,b), \\ 0, & \text{if } t_0 \text{ is outside } (a,b). \end{cases} \tag{15.29}$$

Relations such as

$$\delta(t-t_0) = \frac{du(t-t_0)}{dt}, \qquad \mathcal{L}\{\,\delta(t-t_0)\} = e^{-t_0 p} \qquad (t_0 \geq 0) \tag{15.30}$$

are investigated generally in *distribution theory* (see 12.9.5.3, p. 699).

4. Piecewise Differentiable Functions

The transform of a piecewise differentiable function can be determined easily with the help of the δ function: If $f(t)$ is piecewise differentiable and at the points t_ν $(\nu = 1, 2, \dots, n)$ it has jumps a_ν, then its first derivative can be represented in the form

$$\frac{df(t)}{dt} = f_s'(t) + a_1\delta(t-t_1) + a_2\delta(t-t_2) + \cdots + a_n\delta(t-t_n) \tag{15.31}$$

where $f_s'(t)$ is the usual derivative of $f(t)$, where it is differentiable.
If jumps occur first in the derivative, then similar formulas are valid. In this way, one can easily determine the transform of functions which correspond to curves composed of parabolic arcs of arbitrarily high degree, e.g., curves found empirically. In formal application of (15.13), the values $f(+0), f'(+0), \dots$ should be replaced by zero in the case of a jump.

■ **A:**
$f(t) = \begin{cases} at+b \text{ for } 0<t<t_0, \\ 0 \qquad\quad \text{otherwise,} \end{cases}$ **(Fig. 15.12)**; $f'(t) = a\,u_{t_0}(t) + b\,\delta(t) - (at_0+b)\,\delta(t-t_0)$; $\mathcal{L}\{f'(t)\} = \frac{a}{p}(1-e^{-t_0p}) + b - (at_0+b)\,e^{-t_0p}$; $\mathcal{L}\{f(t)\} = \frac{1}{p}\left[\frac{a}{p} + b - e^{-t_0p}\left(\frac{a}{p} + at_0 + b\right)\right]$.

■ **B:**
$f(t) = \begin{cases} t & \text{for } 0<t<t_0, \\ 2t_0 - t & \text{for } t_0<t<2t_0, \\ 0 & \text{for } t>2t_0, \end{cases}$ **(Fig. 15.13)**; $f'(t) = \begin{cases} 1 & \text{for } 0<t<t_0, \\ -1 & \text{for } t_0<t<2t_0, \\ 0 & \text{for } t>2t_0, \end{cases}$ **(Fig. 15.14)**;

$f''(t) = \delta(t) - \delta(t-t_0) - \delta(t-t_0) + \delta(t-2t_0)$; $\mathcal{L}\{f''(t)\} = 1 - 2e^{-t_0p} + e^{-2t_0p}$; $\mathcal{L}\{f(t)\} = \frac{(\,1-e^{-t_0p}\,)^2}{p^2}$.

■ **C:** $f(t) = \begin{cases} E\,t/t_0 & \text{for} & 0<t<t_0, \\ E & \text{for} & t_0<t<T-t_0, \\ -E(t-T)/t_0 & \text{for} & T-t_0<t<T, \\ 0 & \text{otherwise,} \end{cases}$ **(Fig. 15.15)**;

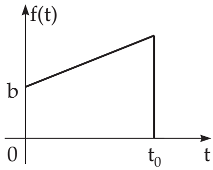


Figure 15.12

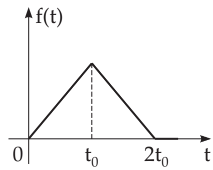


Figure 15.13

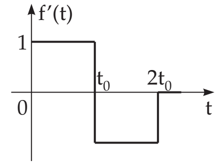


Figure 15.14

$$f'(t) = \begin{cases} E/t_0 & \text{for} \quad 0 < t < t_0, \\ 0 & \text{for} \quad t_0 < t < T - t_0,\ (t > T), \\ -E/t_0 & \text{for} \quad T - t_0 < t < T, \\ 0 & \text{otherwise}, \end{cases} \quad \textbf{(Fig. 15.16)};$$

$$f''(t) = \frac{E}{t_0}\delta(t) - \frac{E}{t_0}\delta(t-t_0) - \frac{E}{t_0}\delta(t-T+t_0) + \frac{E}{t_0}\delta(t-T);\ \mathcal{L}\{f''(t)\} = \frac{E}{t_0}\left[1 - e^{-t_0 p} - e^{-(T-t_0)p} + e^{-Tp}\right];$$

$$\mathcal{L}\{f(t)\} = \frac{E}{t_0}\frac{(1 - e^{-t_0 p})(1 - e^{-(T-t_0)p})}{p^2}.$$

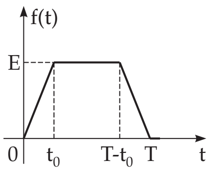


Figure 15.15

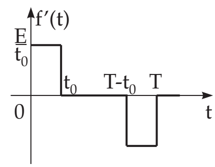


Figure 15.16

■ **D:**

$$f(t) = \begin{cases} t - t^2 & \text{for } 0 < t < 1, \\ 0 & \text{otherwise}, \end{cases} \quad \textbf{(Fig. 15.17)}; \quad f'(t) = \begin{cases} 1 - 2t & \text{for } 0 < t < 1, \\ 0 & \text{otherwise}, \end{cases} \quad \textbf{(Fig. 15.18)};$$

$$f''(t) = -2u_1(t) + \delta(t) + \delta(t-1);$$

$$\mathcal{L}\{f''(t)\} = -\frac{2}{p}(1 - e^{-p}) + 1 + e^{-p}; \quad \mathcal{L}\{f(t)\} = \frac{1 + e^{-p}}{p^2} - \frac{2\,(1 - e^{-p})}{p^3}.$$

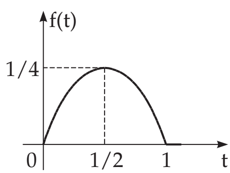


Figure 15.17

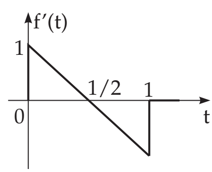


Figure 15.18

5. Periodic Functions

The transform of a periodic function $f^*(t)$ with period T, which is a periodic continuation of a function $f(t)$, can be obtained from the Laplace transform of $f(t)$ multiplied by the *periodization factor*

$$(1 - e^{-Tp})^{-1}. \tag{15.32}$$

■ **A:** The periodic continuation of $f(t)$ from example **B** (see above) with period $T = 2t_0$ is $f^*(t)$ with $\mathcal{L}\{f^*(t)\} = \dfrac{(1-e^{-t_0p})^2}{p^2} \cdot \dfrac{1}{1-e^{-2t_0p}} = \dfrac{1-e^{-t_0p}}{p^2(1+e^{-t_0p})}$.

■ **B:** The periodic continuation of $f(t)$ from example **C** (see above) with period T is $f^*(t)$ with $\mathcal{L}\{f^*(t)\} = \dfrac{E\,(1-e^{-t_0p})\,(1-e^{-(T-t_0)p})}{t_0\,p^2\,(1-e^{-Tp})}$.

15.2.1.4 Dirac δ Function and Distributions

In describing certain technical systems by linear differential equations, functions $u(t)$ and $\delta(t)$ often occur as perturbation or input functions, although the conditions required in 15.2.1.1, **1.** p. 770, are not satisfied: $u(t)$ is discontinuous, and $\delta(t)$ cannot be defined in the sense of classical analysis.
Distribution theory offers a solution by introducing so-called *generalized functions* (*distributions*), so that with the known continuous real functions $\delta(t)$ can also be examined, where the necessary differentiability is also guaranteed. Distributions can be represented in different ways. One of the best known representations is the continuous real linear form, introduced by L. Schwartz (see 12.9.5, p. 698).
Fourier coefficients and Fourier series can be associated uniquely to periodic distributions, analogously to real functions (see 7.4, p. 474).

1. Approximations of the δ Function

Analogously to (15.28), the impulse function $\delta(t)$ can be approximated by a rectangular impulse of width ε and height $1/\varepsilon$ $(\varepsilon > 0)$:

$$f(t,\varepsilon) = \begin{cases} 1/\varepsilon & \text{for } |t| < \varepsilon/2, \\ 0 & \text{for } |t| \ge \varepsilon/2. \end{cases} \qquad (15.33a)$$

Further examples of the approximation of $\delta(t)$ are the error curve (see 2.6.3, p. 73) and Lorentz function (see 2.11.2, p. 95):

$$f(t,\varepsilon) = \frac{1}{\varepsilon\sqrt{2\pi}} e^{-\frac{t^2}{2\varepsilon^2}} \quad (\varepsilon > 0), \qquad (15.33b)$$

$$f(t,\varepsilon) = \frac{\varepsilon/\pi}{t^2+\varepsilon^2} \quad (\varepsilon > 0). \qquad (15.33c)$$

These functions have the common properties:

1. $$\int_{-\infty}^{\infty} f(t,\varepsilon)\,dt = 1. \qquad (15.34a)$$

2. $f(-t,\varepsilon) = f(t,\varepsilon)$, i.e., they are even functions. (15.34b)

3. $$\lim_{\varepsilon\to 0} f(t,\varepsilon) = \begin{cases} \infty & \text{for } t = 0, \\ 0 & \text{for } t \ne 0. \end{cases} \qquad (15.34c)$$

2. Properties of the δ Function

Important properties of the δ function are:

1. $$\int_{x-a}^{x+a} f(t)\delta(x-t)\,dt = f(x) \quad (f \text{ is continuous, } a > 0). \qquad (15.35)$$

2. $$\delta(\alpha x) = \frac{1}{\alpha}\delta(x) \quad (\alpha > 0). \qquad (15.36)$$

3. $$\delta(g(x)) = \sum_{i=1}^{n} \frac{1}{|g'(x_i)|}\delta(x-x_i) \quad \text{with } g(x_i) = 0 \text{ and } g'(x_i) \ne 0 \ (i = 1, 2, \ldots, n). \qquad (15.37)$$

Here all roots of $g(x)$ are considered and they must be simple.

4. *n*-th Derivative of the *δ* Function: After n repeated partial integrations of

$$f^{(n)}(x) = \int_{x-a}^{x+a} f^{(n)}(t)\,\delta(x-t)\,dt, \tag{15.38a}$$

a rule is obtained for the n-th derivative of the δ function:

$$(-1)^n f^{(n)}(x) = \int_{x-a}^{x+a} f(t)\,\delta^{(n)}(x-t)\,dt. \tag{15.38b}$$

15.2.2 Inverse Transformation into the Original Space

To perform an inverse transformation, there are the following possibilities:

1. Using a table of correspondences, i.e., a table with the corresponding original functions and transforms (see **Table 21.13**, p. 1109).

2. Reducing to known correspondences by using some properties of the transformation (see 15.2.2.2, p. 778, and 15.2.2.3, p. 779).

3. Evaluating the inverse formula (see 15.2.2.4, p. 780).

15.2.2.1 Inverse Transformation with the Help of Tables

The use of a table is shown here by an example with **Table 21.13**, p. 1109.
Further tables can be found, e.g., in [15.3].

■ $F(p) = \dfrac{1}{(p^2+\omega^2)(p+c)} = F_1(p)\cdot F_2(p)$, $\mathcal{L}^{-1}\{F_1(p)\} = \mathcal{L}^{-1}\left\{\dfrac{1}{p^2+\omega^2}\right\} = \dfrac{1}{\omega}\sin\omega t = f_1(t)$,
$\mathcal{L}^{-1}\{F_2(p)\} = \mathcal{L}^{-1}\left\{\dfrac{1}{p+c}\right\} = e^{-ct} = f_2(t)$. Applying the convolution theorem (15.23) yields:

$$\begin{aligned} f(t) &= \mathcal{L}^{-1}\{F_1(p)\cdot F_2(p)\} \\ &= \int_0^t f_1(\tau)\cdot f_2(t-\tau)\,d\tau = \int_0^t e^{-c(t-\tau)}\frac{\sin\omega\tau}{\omega}\,d\tau = \frac{1}{c^2+\omega^2}\left(\frac{c\sin\omega t-\omega\cos\omega t}{\omega}+e^{-ct}\right). \end{aligned}$$

15.2.2.2 Partial Fraction Decomposition

1. Principle

In many applications, there are transforms in the form $F(p) = H(p)/G(p)$, where $G(p)$ is a polynomial of p. If the original functions for $H(p)$ and $1/G(p)$ are already known, then the required original function of $F(p)$ can be got by applying the convolution theorem.

2. Simple Real Roots of $G(p)$

If the transform $1/G(p)$ has only simple poles p_ν $(\nu = 1, 2, \ldots, n)$, then it has the following partial fraction decomposition:

$$\frac{1}{G(p)} = \sum_{\nu=1}^{n} \frac{1}{G'(p_\nu)(p-p_\nu)}. \tag{15.39}$$

The corresponding original function is

$$q(t) = \mathcal{L}^{-1}\left\{\frac{1}{G(p)}\right\} = \sum_{\nu=1}^{n} \frac{1}{G'(p_\nu)} e^{p_\nu t}. \tag{15.40}$$

3. The Heaviside Expansion Theorem

If the numerator $H(p)$ is also a polynomial of p with a lower degree than $G(p)$, then we can obtain the original function of $F(p)$ with the help of the Heaviside formula

$$f(t) = \sum_{\nu=1}^{n} \frac{H(p_\nu)}{G'(p_\nu)} e^{p_\nu t}. \tag{15.41}$$

4. Complex Roots

Even in cases when the denominator has simple complex roots, the Heaviside expansion theorem can be used in the same way. The terms belonging to complex conjugate roots can be collected into one quadratic expression, whose inverse transformation can be found in tables also in the case of roots of higher multiplicity.

■ $F(p) = \dfrac{1}{(p+c)(p^2+\omega^2)}$, i.e., $H(p) = 1$, $G(p) = (p+c)(p^2+\omega^2)$, $G'(p) = 3p^2 + 2pc + \omega^2$. The zeroes of $G(p)$ $p_1 = -c,\ p_2 = \mathrm{i}\omega,\ p_3 = -\mathrm{i}\omega$ are all simple. According to the Heaviside theorem one gets $f(t) = \dfrac{1}{\omega^2+c^2}e^{-ct} - \dfrac{1}{2\omega(\omega - \mathrm{i}c)}e^{\mathrm{i}\omega t} - \dfrac{1}{2\omega(\omega+\mathrm{i}c)}e^{-\mathrm{i}\omega t}$ or by using partial fraction decomposition and the table $F(p) = \dfrac{1}{\omega^2+c^2}\left[\dfrac{1}{p+c} + \dfrac{c-p}{p^2+\omega^2}\right]$, $f(t) = \dfrac{1}{\omega^2+c^2}\left[e^{-ct} + \dfrac{c}{\omega}\sin\omega t - \cos\omega t\right]$. These expressions for $f(t)$ are identical.

15.2.2.3 Series Expansion

In order to obtain $f(t)$ from $F(p)$ one can try to expand $F(p)$ into a series $F(p) = \sum\limits_{n=0}^{\infty} F_n(p)$, whose terms $F_n(p)$ are transforms of known functions, i.e., $F_n(p) = \mathcal{L}\{f_n(t)\}$.

1. *F*(*p*) is an Absolutely Convergent Series

If $F(p)$ has an absolutely convergent series

$$F(p) = \sum_{n=0}^{\infty} \frac{a_n}{p^{\lambda_n}}, \tag{15.42}$$

for $|p| > R$, where the values λ_n form an arbitrary increasing sequences of numbers $0 < \lambda_0 < \lambda_1 < \cdots < \lambda_n < \cdots < \cdots \to \infty$, then a termwise inverse transformation is possible:

$$f(t) = \sum_{n=0}^{\infty} a_n \frac{t^{\lambda_n - 1}}{\Gamma(\lambda_n)}. \tag{15.43}$$

Γ denotes the gamma function (see 8.2.5, **6.**, p. 514). In particular, for $\lambda_n = n+1$, i.e., for $F(p) = \sum\limits_{n=0}^{\infty} \dfrac{a_{n+1}}{p^{n+1}}$ the series $f(t) = \sum\limits_{n=0}^{\infty} \dfrac{a_{n+1}}{n\,!}t^n$ is obtained, which is convergent for every real and complex t. Furthermore, one can have an estimation in the form $|f(t)| < C\,e^{c|t|}$ (C, c real constants).

■ $F(p) = \dfrac{1}{\sqrt{1+p^2}} = \dfrac{1}{p}\left(1 + \dfrac{1}{p^2}\right)^{-1/2} = \sum\limits_{n=0}^{\infty} \begin{pmatrix} -\frac{1}{2} \\ n \end{pmatrix} \dfrac{1}{p^{2n+1}}$. After a termwise transformation into the original space the result is $f(t) = \sum\limits_{n=0}^{\infty} \begin{pmatrix} -\frac{1}{2} \\ n \end{pmatrix} \dfrac{t^{2n}}{(2n)\,!} = \sum\limits_{n=0}^{\infty} \dfrac{(-1)^n}{(n\,!)^2}\left(\dfrac{t}{2}\right)^{2n} = J_0(t)$ (Bessel function of 0 order).

2. *F*(*p*) is a Meromorphic Function

If $F(p)$ is a *meromorphic function*, which can be represented as the quotient of two integer functions (of two functions having everywhere convergent power series expansions) which do not have common

roots, and so can be rewritten as the sum of an integer function and infinitely many partial fractions, then the equality

$$\frac{1}{2\pi \mathrm{i}} \int_{c-\mathrm{i}y_n}^{c+\mathrm{i}y_n} e^{tp}\, F(p)\, dp = \sum_{\nu=1}^{n} b_\nu e^{p_\nu t} - \frac{1}{2\pi \mathrm{i}} \int_{(K_n)} e^{tp} F(p)\, dp \tag{15.44}$$

is obtained. Here p_ν ($\nu = 1, 2, \ldots, n$) are the first-order poles of the function $F(p)$, b_ν are the corresponding residues (see 14.3.5.4, p. 753), y_ν are certain values and K_ν are certain curves, for example, half circles in the sense represented in **Fig. 15.19**. The solution $f(t)$ has the form

$$f(t) = \sum_{\nu=1}^{\infty} b_\nu e^{p_\nu t}, \qquad \text{if} \qquad \frac{1}{2\pi \mathrm{i}} \int_{(K_n)} e^{tp} F(p)\, dp \;\to\; 0 \tag{15.45}$$

as $y \to \infty$, what is often not easy to verify.

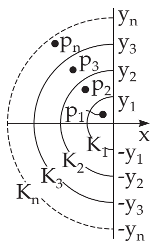


Figure 15.19

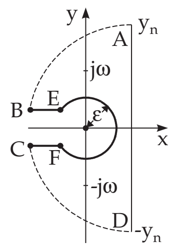


Figure 15.20

In certain cases, e.g., when the rational part of the meromorphic function $F(p)$ is identically zero, the above result is a formal application of the Heaviside expansion theorem to meromorphic functions.

15.2.2.4 Inverse Integral

The inverse formula

$$f(t) = \lim_{y_n \to \infty} \frac{1}{2\pi \mathrm{i}} \int_{c-\mathrm{i}y_n}^{c+\mathrm{i}y_n} e^{tp}\, F(p)\, dp \tag{15.46}$$

represents a complex integral of a function analytic in a certain domain. The usual methods of integration for complex functions can be used, e.g., the residue calculation or certain changes of the path of integration according to the Cauchy integral theorem.

■ $F(p) = \dfrac{p}{p^2+\omega^2} e^{-\sqrt{p}\alpha}$ is double valued because of $\sqrt{p}$. Therefore, we chose the following path of integration **(Fig. 15.20)**: $\dfrac{1}{2\pi \mathrm{i}} \oint_{(K)} e^{tp} \dfrac{p}{p^2+\omega^2} e^{-\sqrt{p}\alpha}\, dp = \int_{\overset{\frown}{AB}} \cdots + \int_{\overset{\frown}{CD}} \cdots + \int_{\overset{\frown}{EF}} \cdots + \int_{\overline{DA}} \cdots + \int_{\overline{BE}} \cdots + \int_{\overline{FC}} \cdots =$ $\sum \operatorname{Res} e^{tp} F(p) = e^{-\alpha\sqrt{\omega/2}} \cos(\omega t - \alpha\sqrt{\omega/2})$. According to the Jordan lemma (see 14.4.3, p. 755), the integral part over $\overset{\frown}{AB}$ and $\overset{\frown}{CD}$ vanishes as $y_n \to \infty$. The integrand remains bounded on the circular arc $\overset{\frown}{EF}$ (radius ε), and the length of the path of integration tends to zero for $\varepsilon \to 0$; so this term of the integral also vanishes. There are to investigate the integrals on the two horizontal segments $\overline{BE}$ and $\overline{FC}$, where it is to consider the upper side ($p = re^{\mathrm{i}\pi}$) and the lower side ($p = re^{-\mathrm{i}\pi}$) of the negative real axis:

$$\int_{-\infty}^{0} F(p)e^{tp}\,dp = -\int_0^\infty e^{-tr}\frac{r}{r^2+\omega^2}e^{-\mathrm{i}\alpha\sqrt{r}}\,dr, \qquad \int_0^{-\infty} F(p)e^{tp}\,dp = \int_0^\infty e^{-tr}\frac{r}{r^2+\omega^2}e^{\,\mathrm{i}\alpha\sqrt{r}}\,dr.$$

Finally one gets:

$$f(t) = e^{-\alpha\sqrt{\omega/2}}\cos\left(\omega t - \alpha\sqrt{\frac{\omega}{2}}\right) - \tfrac{1}{\pi}\int_0^\infty e^{-tr}\frac{r\sin\alpha\sqrt{r}}{r^2+\omega^2}\,dr.$$

15.2.3 Solution of Differential Equations using Laplace Transformation

It has been noticed already from the rules of calculation of the Laplace transformation (see 15.2.1.2, p. 771), that complicated operations, such as differentiation or integration in the original space, can be replaced by simple algebraic operations in the image space using the Laplace transform. Here, some additional conditions are considered, such as initial conditions in using the differentiation rule. These conditions are necessary for the solution of differential equations.

15.2.3.1 Ordinary Linear Differential Equations with Constant Coefficients

1. Principle

The n-th order differential equation of the form

$$y^{(n)}(t) + c_{n-1}\,y^{(n-1)}(t) + \cdots + c_1\,y'(t) + c_0\,y(t) = f(t) \tag{15.47a}$$

with the initial values $y(+0) = y_0$, $y'(+0) = y_0', \ldots, y^{(n-1)}(+0) = y_0^{(n-1)}$ can be transformed by Laplace transformation into the equation

$$\sum_{k=0}^{n} c_k p^k Y(p) - \sum_{k=1}^{n} c_k \sum_{\nu=0}^{k-1} p^{k-\nu-1} y_0^{(\nu)} = F(p) \qquad (c_n = 1). \tag{15.47b}$$

Here $G(p) = \sum\limits_{k=0}^{n} c_k p^k = 0$ is the characteristic equation of the differential equation (see 4.6.2.1, p. 315).

2. First-Order Differential Equations

The original and the transformed equations are:

$$y'(t) + c_0 y(t) = f(t), \quad y(+0) = y_0, \qquad (15.48a) \qquad (p + c_0)\,Y(p) - y_0 = F(p), \qquad (15.48b)$$

where $c_0 = $ const. The solution for $Y(p)$ results in

$$Y(p) = \frac{F(p) + y_0}{p + c_0}. \tag{15.48c}$$

Special case: For $f(t) = \lambda\, e^{\mu t}$ with $F(p) = \dfrac{\lambda}{p-\mu}$, $(\lambda, \mu \text{ const})$: (15.49a)

$$Y(p) = \frac{\lambda}{(p-\mu)(p+c_0)} + \frac{y_0}{p+c_0}, \tag{15.49b}$$

$$y(t) = \frac{\lambda}{\mu + c_0}e^{\mu t} + \left(y_0 - \frac{\lambda}{\mu + c_0}\right)e^{-c_0 t}. \tag{15.49c}$$

3. Second-Order Differential Equations

The original and transformed equations are:

$$y''(t) + 2ay'(t) + by(t) = f(t), \qquad y(+0) = y_0, \qquad y'(+0) = y_0'. \tag{15.50a}$$

$$(p^2 + 2ap + b)\,Y(p) - 2ay_0 - (py_0 + y_0') = F(p). \tag{15.50b}$$

The solution for $Y(p)$ results in

$$Y(p) = \frac{F(p) + (2a+p)\,y_0 + y_0'}{p^2 + 2ap + b}. \tag{15.50c}$$

Distinction of Cases:

a) $b < a^2:\quad G(p) = (p-\alpha_1)(p-\alpha_2) \quad (\alpha_1,\ \alpha_2 \text{ real};\ \alpha_1 \neq \alpha_2),$ (15.51a)

$$q(t) = \mathcal{L}^{-1}\left\{\frac{1}{G(p)}\right\} = \frac{1}{\alpha_1 - \alpha_2}(e^{\alpha_1 t} - e^{\alpha_2 t}). \tag{15.51b}$$

b) $b = a^2:\quad G(p) = (p-\alpha)^2,$ (15.52a) $\qquad q(t) = t\,e^{\alpha t}.$ (15.52b)

c) $b > a^2:\quad G(p)$ has complex roots, (15.53a)

$$q(t) = \mathcal{L}^{-1}\left\{\frac{1}{G(p)}\right\} = \frac{1}{\sqrt{b-a^2}}\,e^{-at}\sin\sqrt{b-a^2}\,t. \tag{15.53b}$$

The solution $y(t)$ can be obtained as the convolution of the original function of the numerator of $Y(p)$ and $q(t)$. The application of the convolution can be avoided if a direct transformation of the right-hand side can be found.

■ The transformed equation for the differential equation $y''(t) + 2y'(t) + 10y(t) = 37\cos 3t + 9e^{-t}$ with $y_0 = 1$ and $y_0' = 0$ is $Y(p) = \dfrac{p+2}{p^2+2p+10} + \dfrac{37p}{(p^2+9)(p^2+2p+10)} + \dfrac{9}{(p+1)(p^2+2p+10)}$. The representation $Y(p) = \dfrac{-p}{p^2+2p+10} - \dfrac{19}{(p^2+2p+10)} + \dfrac{p}{(p^2+9)} + \dfrac{18}{(p^2+9)} + \dfrac{1}{(p+1)}$ follows from partial fraction decomposition of the second and third terms of the right-hand side but not separating the second-order terms into linear ones. The solution after termwise transformation is (see **Table 21.13**, p. 1109) $y(t) = (-\cos 3t - 6\sin 3t)e^{-t} + \cos 3t + 6\sin 3t + e^{-t}$.

4. *n*-th Order Differential Equations

The characteristic equation $G(p) = 0$ of this differential equation (see (15.47a)) has only simple roots $\alpha_1, \alpha_2, \ldots, \alpha_n$, and none of them is equal to zero. Two cases are distinguished for the perturbation function $f(t)$.

1. If the perturbation function $f(t)$ is the jump function $u(t)$ which often occurs in practical problems, then the solution is:

$$u(t) = \begin{cases} 1 & \text{for } t > 0, \\ 0 & \text{for } t < 0, \end{cases} \tag{15.54a}$$

$$y(t) = \frac{1}{G(0)} + \sum_{\nu=1}^{n} \frac{1}{\alpha_\nu G'(\alpha_\nu)} e^{\alpha_\nu t}. \tag{15.54b}$$

2. For a general perturbation function $f(t)$, one gets the solution $\tilde{y}(t)$ from (15.54b) in the form of the Duhamel formula which uses the convolution (see 15.2.1.2, **11.**, p. 773):

$$\tilde{y}(t) = \frac{d}{dt}\int_0^t y(t-\tau)f(\tau)\,d\tau = \frac{d}{dt}[y * f]. \tag{15.55}$$

15.2.3.2 Ordinary Linear Differential Equations with Coefficients Depending on the Variable

Differential equations whose coefficients are polynomials in t can also be solved by Laplace transformation. Applying (15.16), in the image space yields a differential equation, whose order can be lower than the original one.

If the coefficients are first-order polynomials, then the differential equation in the image space is a first-order differential equation and may be it can be solved more easily.

■ Bessel differential equation of 0 order: $t\frac{d^2f}{dt^2}+\frac{df}{dt}+tf=0$ (see (9.52a, p. 562) for $n=0$). The transformation into the image space results in

$$-\frac{d}{dp}[p^2F(p)-pf(0)-f'(0)]+pF(p)-f(0)-\frac{dF(p)}{dp}=0 \quad \text{or} \quad \frac{dF}{dp}=-\frac{p}{p^2+1}F(p)\,.$$

Separation of the variables and integration yields $\log F(p)=-\int\frac{p\,dp}{p^2+1}=-\log\sqrt{p^2+1}+\log C$,

$F(p)=\frac{C}{\sqrt{p^2+1}}$ (C is the integration constant), $f(t)=CJ_0(t)$ (see ■ in 15.2.2.3,**1.**, p. 779 with the Bessel function of 0 order).

15.2.3.3 Partial Differential Equations

1. General Introduction

The solution of a partial differential equation is a function of at least two variables: $u=u(x,t)$. Since the Laplace transformation represents an integration with respect to only one variable, the other variable should be considered as a constant in the transformation:

$$\mathcal{L}\{u(x,t)\}=\int_0^\infty e^{-pt}u(x,t)\,dt=U(x,p). \tag{15.56}$$

x also remains fixed in the transformation of derivatives:

$$\begin{aligned}\mathcal{L}\left\{\frac{\partial u(x,t)}{\partial t}\right\}&=p\,\mathcal{L}\{u(x,t)\}-u(x,+0),\\ \mathcal{L}\left\{\frac{\partial^2 u(x,t)}{\partial t^2}\right\}&=p^2\mathcal{L}\{u(x,t)\}-u(x,+0)p-u_t(x,+0).\end{aligned} \tag{15.57}$$

The differentiation with respect to x is supposed to be interchangeable with the Laplace integral:

$$\mathcal{L}\left\{\frac{\partial u(x,t)}{\partial x}\right\}=\frac{\partial}{\partial x}\mathcal{L}\{u(x,t)\}=\frac{\partial}{\partial x}U(x,p). \tag{15.58}$$

In this way, an ordinary differential equation is obtained in the image space. Furthermore, the boundary and initial conditions are to be transformed into the image space.

2. Solution of the One-Dimensional Heat Conduction Equation for a Homogeneous Medium

1. Formulation of the Problem Suppose the one-dimensional heat conduction equation with vanishing perturbation and for a homogeneous medium is given in the form

$$u_{xx}-a^{-2}u_t=u_{xx}-u_y=0 \tag{15.59a}$$

in the original space $0<t<\infty$, $0<x<l$ and with the initial and boundary conditions

$$u(x,+0)=u_0(x), \quad u(+0,t)=a_0(t), \quad u(l-0,t)=a_1(t). \tag{15.59b}$$

The time coordinate is replaced by $y=at$. (15.59a) is also a parabolic type equation, just as the three-dimensional heat conduction equation (see 9.2.3.3, p. 591).

2. Laplace Transformation The transformed equation is

$$\frac{d^2U}{dx^2}=pU-u_0(x), \tag{15.60a}$$

and the boundary conditions are

$$U(+0,p)=A_0(p), \quad U(l-0,p)=A_1(p). \tag{15.60b}$$

The solution of the transformed equation for zero starting temperature $u_0(x)=0$ is

$$U(x,p)=c_1e^{x\sqrt{p}}+c_2e^{-x\sqrt{p}}. \tag{15.60c}$$

It is a good idea to produce two particular solutions U_1 and U_2 with the properties

$$U_1(0,p)=1,\quad U_1(l,p)=0, \qquad (15.61a) \qquad U_2(0,p)=0,\quad U_2(l,p)=1,\ \text{i.e.,} \qquad (15.61b)$$

$$U_1(x,p)=\frac{e^{(l-x)\sqrt{p}}-e^{-(l-x)\sqrt{p}}}{e^{l\sqrt{p}}-e^{-l\sqrt{p}}}, \qquad (15.61c) \qquad U_2(x,p)=\frac{e^{x\sqrt{p}}-e^{-x\sqrt{p}}}{e^{l\sqrt{p}}-e^{-l\sqrt{p}}}. \qquad (15.61d)$$

The required solution of the transformed equation has the form

$$U(x,p)=A_0(p)\,U_1(x,p)+A_1(p)\,U_2(x,p). \qquad (15.62)$$

3. Inverse Transformation The inverse transformation is especially easy in the case of $l\to\infty$:

$$U(x,p)=a_0(p)e^{-x\sqrt{p}}, \qquad (15.63a) \qquad u(x,t)=\frac{x}{2\sqrt{\pi}}\int\limits_0^t\frac{a_0(t-\tau)}{\tau^{3/2}}\exp\left(-\frac{x^2}{4\tau}\right)d\tau. \qquad (15.63b)$$

15.3 Fourier Transformation

15.3.1 Properties of the Fourier Transformation

15.3.1.1 Fourier Integral

1. Fourier Integral in Complex Representation

The basis of the Fourier transformation is the Fourier integral, also called the *integral formula of Fourier*. If a non-periodic function $f(t)$ satisfies the Dirichlet conditions (see 7.4.1.2, **3.**, p. 475) in an arbitrary finite interval, and furthermore the integral

$$\int\limits_{-\infty}^{+\infty}|f(t)|\,dt \qquad (15.64a)$$ is convergent, then $$f(t)=\frac{1}{2\pi}\int\limits_{-\infty}^{+\infty}\int\limits_{-\infty}^{+\infty}e^{i\omega(t-\tau)}f(\tau)\,d\omega\,d\tau \qquad (15.64b)$$

at every point where the function $f(t)$ is continuous, and

$$\frac{f(t+0)+f(t-0)}{2}=\frac{1}{\pi}\int\limits_0^{\infty}d\omega\int\limits_{-\infty}^{+\infty}f(\tau)\cos\omega\,(t-\tau)\,d\tau \qquad (15.64c)$$

at the points of discontinuity.

2. Equivalent Representations

Other equivalent forms for the Fourier integral (15.64b) are:

1. $$f(t)=\frac{1}{2\pi}\int\limits_{-\infty}^{+\infty}\int\limits_{-\infty}^{+\infty}f(\tau)\cos\left[\,\omega\,(t-\tau)\,\right]d\omega\,d\tau. \qquad (15.65a)$$

2. $$f(t)=\int\limits_0^{\infty}\left[\,a(\omega)\cos\omega t+b(\omega)\sin\omega t\,\right]d\omega \qquad (15.65b)$$ with the coefficients

$$a(\omega)=\frac{1}{\pi}\int\limits_{-\infty}^{+\infty}f(t)\cos\omega t\,dt \qquad (15.65c) \qquad b(\omega)=\frac{1}{\pi}\int\limits_{-\infty}^{+\infty}f(t)\sin\omega t\,dt. \qquad (15.65d)$$

3. $$f(t)=\int\limits_0^{\infty}A(\omega)\cos\left[\,\omega t+\psi(\omega)\,\right]d\omega. \qquad (15.66)$$

4. $f(t) = \int_{0}^{\infty} A(\omega) \sin[\omega t + \varphi(\omega)]\, d\omega.$ (15.67)

The following relations are valid here:

$$A(\omega) = \sqrt{a^2(\omega) + b^2(\omega)}\,, \quad (15.68a) \qquad \varphi(\omega) = \psi(\omega) + \frac{\pi}{2}\,, \quad (15.68b)$$

$$\cos\psi(\omega) = \frac{a(\omega)}{A(\omega)}, \quad (15.68c) \qquad \sin\psi(\omega) = \frac{b(\omega)}{A(\omega)}\,, \quad (15.68d)$$

$$\cos\varphi(\omega) = \frac{b(\omega)}{A(\omega)}\,, \quad (15.68e) \qquad \sin\varphi(\omega) = \frac{a(\omega)}{A(\omega)}\,. \quad (15.68f)$$

15.3.1.2 Fourier Transformation and Inverse Transformation

1. Definition of the Fourier Transformation

The Fourier transformation is an integral transformation of the form (15.1a), which comes from the Fourier integral (15.64b) by substituting

$$F(\omega) = \int_{-\infty}^{+\infty} e^{-\mathrm{i}\omega\tau} f(\tau)\, d\tau. \quad (15.69)$$

The following relation is valid between the real original function $f(t)$ and the usually complex transform $F(\omega)$:

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{\mathrm{i}\omega t} F(\omega)\, d\omega. \quad (15.70)$$

In the brief notation one uses $\mathcal{F}$:

$$F(\omega) = \mathcal{F}\{\, f(t)\,\} = \int_{-\infty}^{+\infty} e^{-\mathrm{i}\omega t} f(t)\, dt. \quad (15.71)$$

The original function $f(t)$ is Fourier transformable if the integral (15.69), i.e., an improper integral with the parameter ω, exists. If the Fourier integral does not exist as an ordinary improper integral, then it is considered as the Cauchy principal value (see 8.2.3.3, **1.**, p. 510). The transform $F(\omega)$ is also called the *Fourier transform*; it is bounded, continuous, and it tends to zero for $|\omega| \to \infty$:

$$\lim_{|\omega|\to\infty} F(\omega) = 0. \quad (15.72)$$

The existence and boundedness of $F(\omega)$ follow directly from the obvious inequality

$$|F(\omega)| \le \int_{-\infty}^{+\infty} |e^{-\mathrm{i}\omega t} f(t)|\, dt \le \int_{-\infty}^{+\infty} |f(t)|\, dt. \quad (15.73)$$

The existence of the Fourier transform is a sufficient condition for the continuity of $F(\omega)$ and for the properties $F(\omega) \to 0$ for $|\omega| \to \infty$. This statement is often used in the following form: If the function $f(t)$ in $(-\infty, \infty)$ is absolutely integrable, then its Fourier transform is a continuous function of ω, and (15.72) holds.

The following functions are not Fourier transformable: Constant functions, arbitrary periodic functions (e.g., $\sin\omega t, \cos\omega t$), power functions, polynomials, exponential functions (e.g., $e^{\alpha t}$, hyperbolic functions).

2. Fourier Cosine and Fourier Sine Transformation

In the Fourier transformation (15.71), the integrand can be decomposed into a sine and a cosine part. So, one gets the sine and the cosine Fourier transformation.

1. Fourier Sine Transformation

$$F_s(\omega) = \mathcal{F}_s\{\, f(t)\,\} = \int_0^\infty f(t)\sin(\omega t)\,dt. \tag{15.74a}$$

2. Fourier Cosine Transformation

$$F_c(\omega) = \mathcal{F}_c\{\, f(t)\,\} = \int_0^\infty f(t)\cos(\omega\, t)\,dt. \tag{15.74b}$$

3. Conversion Formulas Between the Fourier sine (15.74a) and the Fourier cosine transformation (15.74b) on one hand, and the Fourier transformation (15.71) on the other hand, the following relations are valid:

$$F(\omega) = \mathcal{F}\{\, f(t)\,\} = \mathcal{F}_c\{\, f(t) + f(-t)\,\} - \mathrm{i}\mathcal{F}_s\{\, f(t) - f(-t)\,\}, \tag{15.75a}$$

$$F_s(\omega) = \frac{\mathrm{i}}{2}\mathcal{F}\{\, f(|t|)\operatorname{sign} t\,\}, \quad (15.75\text{b}) \qquad F_c(\omega) = \frac{1}{2}\mathcal{F}\{\, f(|t|)\,\}. \quad (15.75\text{c})$$

For an even or for an odd function $f(t)$ the following representations hold:

$$\begin{aligned} &f(t)\ \text{even:}\ \mathcal{F}\{\, f(t)\,\} = 2\mathcal{F}_c\{\, f(t)\,\},\\ &f(t)\ \text{odd:}\ \ \mathcal{F}\{\, f(t)\,\} = -2\mathrm{i}\mathcal{F}_s\{\, f(t)\,\}. \end{aligned} \tag{15.75d}$$

3. Exponential Fourier Transformation

Differently from the definition of $F(\omega)$ in (15.71), the transform

$$F_e(\omega) = \mathcal{F}_e\{f(t)\} = \frac{1}{2}\int_{-\infty}^{+\infty} e^{\mathrm{i}\omega t} f(t)\,dt \tag{15.76}$$

is called the *exponential Fourier transformation*, so that

$$F(\omega) = 2F_e(-\omega). \tag{15.77}$$

4. Tables of the Fourier Transformation

Based on formulas (15.75a,b,c) one either does not need special tables for the corresponding Fourier sine and Fourier cosine transformations, or one uses tables for Fourier sine and Fourier cosine transformations and calculates $\mathcal{F}(\omega)$ with the help of (15.75a,b,c). In **Table 21.14.1** (see p. 1114) and **Table 21.14.2** (see p. 1120) the Fourier sine transforms $\mathcal{F}_s(\omega)$, the Fourier cosine transforms $\mathcal{F}_c(\omega)$ respectively, in **Table 21.14.3** (see p. 1125) for some functions the Fourier transform $\mathcal{F}(\omega)$ and in **Table 21.14.4** (see p. 1127) the exponential transform $\mathcal{F}_e(\omega)$ are given.

■ The function of the unipolar rectangular impulse $f(t) = 1$ for $|t| < t_0$, $f(t) = 0$ for $|t| > t_0$ (A.1) **(Fig. 15.21)** satisfies the assumptions of the existence of the Fourier integral (15.64a). According to (15.65c,d) the coefficients are $a(\omega) = \dfrac{1}{\pi}\displaystyle\int_{-t_0}^{+t_0}\cos\omega\, t\,dt = \dfrac{2}{\pi\omega}\sin\omega\, t_0$ and $b(\omega) = \dfrac{1}{\pi}\displaystyle\int_{-t_0}^{+t_0}\sin\omega\, t\,dt = 0$ (A.2) and so from (15.65b) follows $f(t) = \dfrac{2}{\pi}\displaystyle\int_0^\infty \frac{\sin\omega\, t_0\cos\omega\, t}{\omega}\,d\omega$ (A.3).

5. Spectral Interpretation of the Fourier Transformation

Analogously to the Fourier series of a periodic function, the Fourier integral for a non-periodic function has a simple physical interpretation. A function $f(t)$, for which the Fourier integral exists, can be represented according to (15.66) and (15.67) as a sum of sinusoidal vibrations with continuously changing frequency ω in the form

$$A(\omega)\,d\omega\,\sin[\,\omega t+\varphi(\omega)\,], \qquad (15.78a) \qquad\qquad A(\omega)\,d\omega\,\cos[\,\omega t+\psi(\omega)\,]. \qquad (15.78b)$$

The expression $A(\omega)\,d\omega$ gives the amplitude of the wave components and $\varphi(\omega)$ and $\psi(\omega)$ are the phases. The same interpretation holds for the complex formulation: The function $f(t)$ is a sum (or integral) of summands depending on ω of the form

$$\frac{1}{2\pi}F(\omega)\,d\omega\,e^{i\omega t}, \qquad (15.79)$$

where the quantity $\frac{1}{2\pi}F(\omega)$ also determines the amplitude and the phase of all the parts.

This *spectral interpretation* of the Fourier integral and the Fourier transformation has a big advantage in applications in physics and engineering. The transform

$$F(\omega)=|F(\omega)|\,e^{i\psi(\omega)} \quad \text{or} \quad F(\omega)=|F(\omega)|\,e^{i\varphi(\omega)} \qquad (15.80a)$$

is called the *spectrum* or *frequency spectrum* of the function $f(t)$, the quantity

$$|F(\omega)|=\pi\,A(\omega) \qquad (15.80b)$$

is the *amplitude spectrum* and $\varphi(\omega)$ and $\psi(\omega)$ are the *phase spectra* of the function $f(t)$. The relation between the spectrum $F(\omega)$ and the coefficients (15.65c,d) is

$$F(\omega)=\pi[\,a(\omega)-ib(\omega)\,], \qquad (15.81)$$

from which one gets the following statements:

1. If $f(t)$ is a real function, then the amplitude spectrum $|F(\omega)|$ is an even function of ω, and the phase spectrum is an odd function of ω.

2. If $f(t)$ is a real and even function, then its spectrum $F(\omega)$ is real, and if $f(t)$ is real and odd, then the spectrum $F(\omega)$ is imaginary.

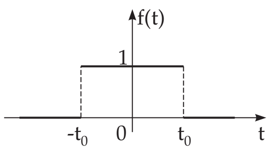


Figure 15.21

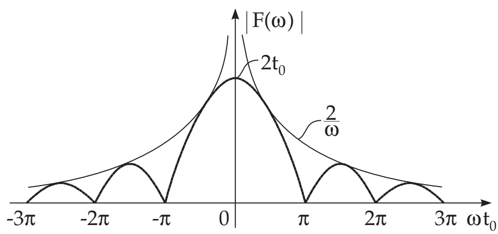


Figure 15.22

■ Substituting the result (A.2) for the unipolar rectangular impulse function on p. 786 into (15.81), then one gets for the transform $F(\omega)$ and for the amplitude spectrum $|F(\omega)|$ **(Fig. 15.22)**
$F(\omega)=\mathcal{F}\{\,f(t)\,\}=\pi a(\omega)=2\frac{\sin\omega t_0}{\omega}$ (A.3), $|F(\omega)|=2\left|\frac{\sin\omega t_0}{\omega}\right|$ (A.4). The points of contact of the amplitude spectrum $|F(\omega)|$ with the hyperbola $\frac{2}{\omega}$ are at $\omega t_0=\pm(2n+1)\frac{\pi}{2}$ $(n=0,1,2,\ldots)$.

15.3.1.3 Rules of Calculation with the Fourier Transformation

As it has been already pointed out for the Laplace transformation, the rules of calculation with integral transformations mean the mappings of certain operations in the original space into operations in the image space. Supposing that both functions $f(t)$ and $g(t)$ are absolutely integrable in the interval $(-\infty,\ \infty)$ and their Fourier transforms are

$$F(\omega)=\mathcal{F}\{\,f(t)\,\} \quad \text{and} \quad G(\omega)=\mathcal{F}\{\,g(t)\,\} \qquad (15.82)$$

then the following rules are valid.

1. Addition or Linearity Laws

If α and β are two coefficients from $(-\infty, \infty)$, then:

$$\mathcal{F}\{\, \alpha f(t) + \beta g(t)\,\} = \alpha F(\omega) + \beta G(\omega). \tag{15.83}$$

2. Similarity Law

For real $\alpha \neq 0$,

$$\mathcal{F}\{\, f(t/\alpha)\,\} = |\alpha|\, F(\alpha\omega). \tag{15.84}$$

3. Shifting Theorem

For real $\alpha \neq 0$ and real β,

$$\mathcal{F}\{\, f(\alpha t + \beta)\,\} = (1/|\alpha|)\, e^{\mathrm{i}\beta\omega/\alpha} F(\omega/\alpha) \quad \text{or} \tag{15.85a}$$

$$\mathcal{F}\{\, f(t - t_0)\,\} = e^{-\mathrm{i}\omega t_0} F(\omega). \tag{15.85b}$$

If t_0 is replaced by $-t_0$ in (15.85b), then

$$\mathcal{F}\{\, f(t + t_0)\,\} = e^{\mathrm{i}\omega t_0} F(\omega). \tag{15.85c}$$

4. Frequency-Shift Theorem

For real $\alpha > 0$ and $\beta \in (-\infty, \infty)$,

$$\mathcal{F}\{\, e^{\mathrm{i}\beta t} f(\alpha t)\,\} = (1/\alpha) F((\omega - \beta)/\alpha) \quad \text{or} \tag{15.86a}$$

$$\mathcal{F}\{\, e^{\mathrm{i}\omega_0 t} f(t)\,\} = F(\omega - \omega_0). \tag{15.86b}$$

5. Differentiation in the Image Space

If the function $t^n f(t)$ is absolutely integrable in $(-\infty, \infty)$, then the Fourier transform of the function $f(t)$ has n continuous derivatives, which can be determined for $k = 1, 2, \ldots, n$ as

$$\frac{d^k F(\omega)}{d\omega^k} = \int\limits_{-\infty}^{+\infty} \frac{\partial^k}{\partial\omega^k} \left[e^{-\mathrm{i}\omega t} f(t) \right] dt = (-1)^k \int\limits_{-\infty}^{+\infty} e^{-\mathrm{i}\omega t} t^k f(t)\, dt, \tag{15.87a}$$

where

$$\lim_{\omega \to \pm\infty} \frac{d^k F(\omega)}{d\omega^k} = 0. \tag{15.87b}$$

With the above assumptions these relations imply that

$$\mathcal{F}\{\, t^n f(t)\,\} = \mathrm{i}^n \frac{d^n F(\omega)}{d\omega^n}. \tag{15.87c}$$

6. Differentiation in the Original Space

1. First Derivative If a function $f(t)$ is continuous and absolutely integrable in $(-\infty, \infty)$ and it tends to zero for $t \to \pm\infty$, and the derivative $f'(t)$ exists everywhere except, maybe, at certain points, and this derivative is absolutely integrable in $(-\infty, \infty)$, then

$$\mathcal{F}\{\, f'(t)\,\} = \mathrm{i}\omega\, \mathcal{F}\{\, f(t)\,\}. \tag{15.88a}$$

2. *n*-th Derivative If the requirements of the theorem for the first derivative are valid for all derivatives up to $f^{(n-1)}$, then

$$\mathcal{F}\{\, f^{(n)}(t)\,\} = (\mathrm{i}\omega)^n \mathcal{F}\{\, f(t)\,\}. \tag{15.88b}$$

These rules of differentiation will be used in the solution of differential equations (see 15.3.2, p. 791).

7. Integration in the Image Space

$$\int\limits_{\alpha_1}^{\alpha_2} F(\omega)\, d\omega = \mathrm{i}[G(\alpha_2) - G(\alpha_1)] \;\text{ with }\; G(\omega) = \mathcal{F}\{g(t)\} \;\text{ and }\; g(t) = \frac{f(t)}{t}\,. \tag{15.89}$$

8. Integration in the Original Space and the Parseval Formula

1. Integration Theorem If the assumption

$$\int_{-\infty}^{+\infty} f(t)\,dt = 0 \quad (15.90a) \qquad \text{is fulfilled, then} \qquad \mathcal{F}\left\{\int_{-\infty}^{t} f(t)\,dt\right\} = \frac{1}{\mathrm{i}\omega}F(\omega). \quad (15.90b)$$

2. Parseval Formula If the function $f(t)$ and its square are integrable in the interval $(-\infty,\infty)$, then

$$\int_{-\infty}^{+\infty} |f(t)|^2\,dt = \frac{1}{2\pi}\int_{-\infty}^{+\infty} |F(\omega)|^2\,d\omega. \quad (15.91)$$

9. Convolution

The *two-sided convolution*

$$f_1(t) * f_2(t) = \int_{-\infty}^{+\infty} f_1(\tau) f_2(t-\tau)\,d\tau \quad (15.92)$$

is considered in the interval $(-\infty,\infty)$ and it exists under the assumptions that the functions $f_1(t)$ and $f_2(t)$ are absolutely integrable in the interval $(-\infty,\infty)$. If $f_1(t)$ and $f_2(t)$ both vanish for $t<0$, then one gets the *one-sided convolution* from (15.92)

$$f_1(t) * f_2(t) = \begin{cases} \displaystyle\int_0^t f_1(\tau) f_2(t-\tau)\,d\tau & \text{for } t \ge 0, \\ 0 & \text{for } t < 0. \end{cases} \quad (15.93)$$

So, it is a special case of the two-sided convolution. While the Fourier transformation uses the two-sided convolution, the Laplace transformation uses the one-sided convolution.

For the Fourier transformation of a two-sided convolution

$$\mathcal{F}\{ f_1(t) * f_2(t) \} = \mathcal{F}\{ f_1(t) \} \cdot \mathcal{F}\{ f_2(t) \} \quad (15.94)$$

holds, if both integrals

$$\int_{-\infty}^{+\infty} |f_1(t)|^2\,dt \qquad \text{and} \qquad \int_{-\infty}^{+\infty} |f_2(t)|^2\,dt \quad (15.95)$$

exist, i.e., the functions and their squares are integrable in the interval $(-\infty,\infty)$.

■ Calculation of the two-sided convolution $\psi(t) = f(t) * f(t) = \int_{-\infty}^{+\infty} f(\tau) f(t-\tau)\,d\tau$ (A.1) for the function of the unipolar rectangular impulse function (A.1) in 15.3.1.2, **4.**, p. 786.

Since $\psi(t) = \int_{-t_0}^{t_0} f(t-\tau)\,d\tau = \int_{t-t_0}^{t+t_0} f(\tau)\,d\tau$ (A.2) one gets for $t < -2t_0$ and $t > 2t_0$, $\psi(t) = 0$ and for $-2t_0 \le t \le 0$, $\psi(t) = \int_{-t_0}^{t+t_0} d\tau = t + 2t_0$. (A.3)

Analogously, for $0 < t \le 2t_0$: $\psi(t) = \int_{t-t_0}^{t_0} d\tau = -t + 2t_0$ (A.4) holds.

Altogether, for this convolution **(Fig. 15.23)**

$$\psi(t) = f(t) * f(t) = \begin{cases} t + 2t_0 & \text{for } -2t_0 \le t \le 0, \\ -t + 2t_0 & \text{for } 0 < t \le 2t_0, \\ 0 & \text{for } |t| > 2t_0 \end{cases} \quad (A.5)$$

follows. For the Fourier transform $F(\omega)$ of the unipolar rectangular impulse (A.1) (see p. 786 and

Fig. 15.21) $\Psi(\omega) = \mathcal{F}\{\,\psi(t)\,\} = \mathcal{F}\{\,f(t)\,*\,f(t)\} = [\,F(\omega)\,]^2 = 4\frac{\sin^2\omega\,t_0}{\omega^2}$ (A.6) follows and for the amplitude spectrum of the function $f(t)$ $|F(\omega)| = 2\left|\frac{\sin\,\omega\,t_0}{\omega}\right|$ (A.7) holds.

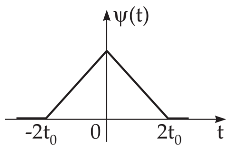


Figure 15.23

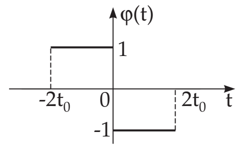


Figure 15.24

10. Comparing the Fourier and Laplace Transformations

There is a strong relation between the Fourier and Laplace transformation, since the Fourier transformation is a special case of the Laplace transformation with $p = \mathrm{i}\omega$. Consequently, every Fourier transformable function is also Laplace transformable, while the reverse statement is not valid for every $f(t)$. **Table 15.2** contains comparisons of several properties of both integral transformations.

Table 15.2 Comparison of the properties of the Fourier and the Laplace transformation

Fourier transformation	**Laplace transformation**
$F(\omega) = \mathcal{F}\{\,f(t)\,\} = \int\limits_{-\infty}^{+\infty} e^{-\mathrm{i}\omega\,t} f(t)\,dt$ ω is real, it has a physical meaning, e.g., frequency.	$F(p) = \mathcal{L}\{\,f(t), p\,\} = \int\limits_{0}^{\infty} e^{-pt} f(t)\,dt$ p is complex, $p = r + \mathrm{i}x$.
One shifting theorem.	Two shifting theorems.
interval: $(-\infty, +\infty)$ Solution of differential equations, problems described by two-sided domain, e.g., the wave equation.	interval: $[\,0, \infty)$ Solution of differential equations, problems described by one-sided domain, e.g., the heat conduction equation.
Differentiation law contains no initial values.	Differentiation law contains initial values.
Convergence of the Fourier integral depends only on $f(t)$.	Convergence of the Laplace integral can be improved by the factor e^{-pt}.
It satisfies the two-sided convolution law.	It satisfies the one-sided convolution law.

15.3.1.4 Transforms of Special Functions

■ **A:** Which image function belongs to the original function $f(t) = e^{-a|t|}$, $\mathrm{Re}\,a > 0$ (A.1)? Considering that $|t| = -t$ for $t < 0$ and $|t| = t$ for $t > 0$ with (15.71) one gets: $\int_{-A}^{+A} e^{-\mathrm{i}\omega\,t - a|t|} dt =$

$$\int_{-A}^{0} e^{-(\mathrm{i}\omega-a)t}\,dt + \int_{0}^{+A} e^{-(\mathrm{i}\omega+a)t}\,dt = -\frac{e^{-(\mathrm{i}\omega-a)t}}{\mathrm{i}\omega - a}\Bigg|_{-A}^{0} - \frac{e^{-(\mathrm{i}\omega+a)t}}{\mathrm{i}\omega + a}\Bigg|_{0}^{+A} = \frac{-1 + e^{(\mathrm{i}\omega-a)A}}{\mathrm{i}\omega - a} + \frac{1 - e^{-(\mathrm{i}\omega+a)A}}{\mathrm{i}\omega + a} \quad \text{(A.2)}.$$

Since $|e^{-aA}| = e^{-A\,\mathrm{Re}\,a}$ and $\mathrm{Re}\,a > 0$, the limit of (A2) exists for $A \to \infty$, so that $F(\omega) = \mathcal{F}\{\,e^{-a|t|}\,\} = \frac{2a}{a^2 + \omega^2}$ (A.3).

■ **B:** Which image function belongs to the original function $f(t) = e^{-at}$, $\mathrm{Re}\,a > 0$? The function is

not Fourier transformable, since the limit of $\int_{-A}^{A} e^{-\mathrm{i}\omega t - at}\,dt$ does not exist for $A \to \infty$.

■ **C:** Determination of the Fourier transform of the bipolar rectangular impulse function **(Fig. 15.24)**

$$\varphi(t) = \begin{cases} 1 & \text{for } -2t_0 < t < 0, \\ -1 & \text{for } 0 < t < 2t_0, \\ 0 & \text{for } |t| > 2t_0\,, \end{cases} \qquad \text{(C.1)}$$

where $\varphi(t)$ can be expressed by using equation (A.1) given for the unipolar rectangular impulse on p. 786. There is $\varphi(t) = f(t+t_0) - f(t-t_0)$ (C.2). With the Fourier transformation according to (15.85b, 15.85c) one gets $\Phi(\omega) = \mathcal{F}\{\,\varphi(t)\,\} = e^{\mathrm{i}\omega\, t_0}F(\omega) - e^{-\mathrm{i}\omega\, t_0}F(\omega)$, (C.3) from which, using (A.1),

$\phi(\omega) = (e^{\mathrm{i}\omega\, t_0} - e^{-\mathrm{i}\omega\, t_0})\dfrac{2\sin\,\omega\, t_0}{\omega} = 4\mathrm{i}\dfrac{\sin^2\omega\, t_0}{\omega}$ (C.4) follows.

■ **D:** Image function of a damped oscillation: The damped oscillation represented in **Fig. 15.25a** is given by the function $f(t) = \begin{cases} 0 & \text{for } t < 0, \\ e^{-\alpha t}\cos\,\omega_0 t & \text{for } t \geq 0. \end{cases}$

To simplify the calculations, the Fourier transformation is calculated with the complex function $f^*(t) = e^{(-\alpha + \mathrm{i}\omega_0)t}$, with $f(t) = \mathrm{Re}\,(f^*(t))$. The Fourier transformation gives

$$\mathcal{F}\{\,f^*(t)\,\} = \int_0^\infty e^{-\mathrm{i}\omega\, t} e^{(-\alpha+\mathrm{i}\omega_0)t}\,dt = \int_0^\infty e^{(-\alpha+(\omega-\omega_0)\mathrm{i}\, t}\,dt = \left.\frac{e^{-\alpha t}e^{\mathrm{i}(\omega-\omega_0)t}}{-\alpha+\mathrm{i}(\omega_0-\omega)}\right|_0^\infty = \frac{1}{\alpha - \mathrm{i}\omega_0 - \omega)} =$$

$\dfrac{\alpha + \mathrm{i}(\omega_0-\omega)}{\alpha^2 + (\omega-\omega_0)^2}$. The result is the Lorentz or Breit-Wigner curve (see also 2.11.2, p. 95)

$\mathcal{F}\{\,f(t)\,\} = \dfrac{\alpha}{\alpha^2 + (\omega-\omega_0)^2}$ **(Fig. 15.25b)**. A damped oscillation in the time domain corresponds to a unique peak in the frequency domain.

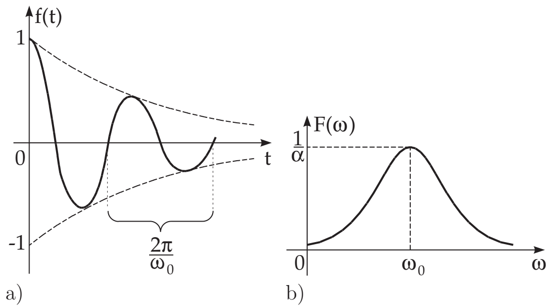


Figure 15.25

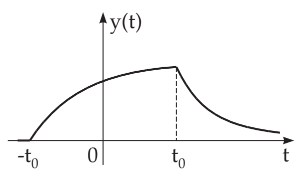


Figure 15.26

15.3.2 Solution of Differential Equations using the Fourier Transformation

Analogously to Laplace transformation, an important field of application of the Fourier transformation is the solution of differential equations, since these equations can be transformed by the integral transformation into a simple form. In the case of ordinary differential equations one gets algebraic equations, in the case of partial differential equations one gets ordinary differential equations.

15.3.2.1 Ordinary Linear Differential Equations

The differential equation

$$y'(t) + a\,y(t) = f(t) \qquad \text{with} \quad f(t) = \begin{cases} 1 & \text{for } |t| < t_0, \\ 0 & \text{for } |t| \geq t_0, \end{cases} \tag{15.96a}$$

i.e., with the function $f(t)$ of **Fig. 15.21**, is transformed by the Fourier transformation

$$\mathcal{F}\{\, y(t) \,\} = Y(\omega) \tag{15.96b}$$

into the algebraic equation

$$\mathrm{i}\omega\, Y + aY = \frac{2\sin\,\omega\, t_0}{\omega}, \quad (15.96\text{c}) \quad \text{giving} \qquad Y(\omega) = 2\,\frac{\sin\,\omega t_0}{\omega(a+\mathrm{i}\omega)}\,. \tag{15.96d}$$

The inverse transformation gives

$$y(t) = \mathcal{F}^{-1}\{\, Y(\omega) \,\} = \mathcal{F}^{-1}\left\{ 2\,\frac{\sin\,\omega t_0}{\omega(a+\mathrm{i}\omega)} \right\} = \frac{1}{\pi}\int\limits_{-\infty}^{+\infty} \frac{e^{\mathrm{i}\omega t}\sin\,\omega t_0}{\omega(a+\mathrm{i}\omega)}\,d\omega \tag{15.96e}$$

and

$$y(t) = \begin{cases} 0 & \text{for } -\infty < t < -t_0, \\ \dfrac{1}{a}\left[1 - e^{-a(t+t_0)}\right] & \text{for } -t_0 \leq t \leq +t_0, \\ \dfrac{1}{a}\left[e^{-a(t-t_0)} - e^{-a(t-t_0)}\right] & \text{for } \quad t_0 < t < \infty\,. \end{cases} \tag{15.96f}$$

Function (15.96f) is represented graphically in **Fig. 15.26**.

15.3.2.2 Partial Differential Equations

1. General Remarks

The solution of a partial differential equation is a function of at least two variables: $u = u(x,t)$. As the Fourier transformation is an integration with respect to only one variable, the other variable is considered a constant during the transformation. Here the variable x is kept constant and the transformation is to be performed with respect to t:

$$\mathcal{F}\{\, u(x,t) \,\} = \int\limits_{-\infty}^{+\infty} e^{-\mathrm{i}\omega\, t} u(x,t)\,dt = U(x,\omega). \tag{15.97}$$

During the transformation of the derivatives the variable x is again kept constant:

$$\mathcal{F}\left\{ \frac{\partial^{(n)} u(x,t)}{\partial t^n} \right\} = (\mathrm{i}\omega)^n \mathcal{F}\{\, u(x,t) \,\} = (\mathrm{i}\omega)^n U(x,\omega). \tag{15.98}$$

The differentiation with respect to x is supposed to be interchangeable with the Fourier integral:

$$\mathcal{F}\left\{ \frac{\partial u(x,t)}{\partial x} \right\} = \frac{\partial}{\partial x}\mathcal{F}\{\, u(x,t) \,\} = \frac{\partial}{\partial x} U(x,\omega)\,. \tag{15.99}$$

In this way an ordinary differential equation is obtained in the image space. Furthermore the boundary and initial conditions are to be transformed into the image space.

2. Solution of the One-Dimensional Wave Equation for a Homogeneous Medium

1. Formulation of the Problem The one-dimensional wave equation with vanishing perturbation term and for a homogeneous medium is:

$$u_{xx} - u_{tt} = 0. \tag{15.100a}$$

Like the three-dimensional wave equation (see 9.2.3.2,p. 590), the equation (15.100a) is a partial differential equation of hyperbolic type. The Cauchy problem is correctly defined by the following initial conditions

$$u(x,0) = f(x) \quad (-\infty < x < \infty), \quad u_t(x,0) = g(x) \quad (0 \le t < \infty). \tag{15.100b}$$

2. Fourier Transformation The Fourier transformation is to be performed with respect to x where the time coordinate is kept constant:

$$\mathcal{F}\{\, u(x,t)\,\} = U(\omega,t). \tag{15.101a}$$

One gets:

$$(\mathrm{i}\omega)^2 U(\omega,t) - \frac{d^2U(\omega,t)}{dt^2} = 0 \qquad \text{with} \tag{15.101b}$$

$$\mathcal{F}\{\, u(x,0)\,\} = U(\omega,0) = \mathcal{F}\{\, f(x)\,\} = F(\omega), \tag{15.101c}$$

$$\mathcal{F}\{\, u_t(x,0)\,\} = U'(\omega,0) = \mathcal{F}\{\, g(x)\,\} = G(\omega). \tag{15.101d}$$

$$\omega^2 U + U'' = 0. \tag{15.101e}$$

The result is an ordinary differential equation with respect to t with the parameter ω of the transform. The general solution of this known differential equation with constant coefficients is

$$U(\omega,t) = C_1 e^{\mathrm{i}\omega t} + C_2 e^{-\mathrm{i}\omega t}. \tag{15.102a}$$

Determining the constants C_1 and C_2 from the initial values

$$U(\omega,0) = C_1 + C_2 = F(\omega)\,, \quad U'(\omega,0) = \mathrm{i}\omega\, C_1 - \mathrm{i}\omega\, C_2 = G(\omega)\,, \tag{15.102b}$$

gives

$$C_1 = \frac{1}{2}[\, F(\omega) + \frac{1}{\mathrm{i}\omega}G(\omega)\,], \quad C_2 = \frac{1}{2}[\, F(\omega) - \frac{1}{\mathrm{i}\omega}G(\omega)\,]. \tag{15.102c}$$

The solution is therefore

$$U(\omega,t) = \frac{1}{2}[\, F(\omega) + \frac{1}{\mathrm{i}\omega}G(\omega)\,]e^{\mathrm{i}\omega t} + \frac{1}{2}[\, F(\omega) - \frac{1}{\mathrm{i}\omega}G(\omega)\,]e^{-\mathrm{i}\omega t}. \tag{15.102d}$$

3. Inverse Transformation Using the shifting theorem

$$\mathcal{F}\{\, f(ax+b)\,\} = 1/a \cdot e^{\mathrm{i}b\omega/a}F(\omega/a), \tag{15.103a}$$

for the inverse transformation of $F(\omega)$, yields

$$\mathcal{F}^{-1}\{\, e^{\mathrm{i}\omega t}F(\omega)\,\} = f(x+t), \quad \mathcal{F}^{-1}[\, e^{-\mathrm{i}\omega t}F(\omega)\,] = f(x-t)\,. \tag{15.103b}$$

Applying the integration rule

$$\mathcal{F}\left\{\int_{-\infty}^{x} f(\tau)\, d\tau\right\} = \frac{1}{\mathrm{i}\omega}F(\omega) \quad \text{gives} \tag{15.103c}$$

$$\mathcal{F}^{-1}\left\{\frac{1}{\mathrm{i}\omega}G(\omega)e^{\mathrm{i}\omega t}\right\} = \int_{-\infty}^{x} \mathcal{F}^{-1}\{\, G(\omega)e^{\mathrm{i}\omega t}\}\, d\tau = \int_{-\infty}^{x} g\,(\tau+t)\, d\tau = \int_{-\infty}^{x+t} g\,(z)\, dz \tag{15.103d}$$

after substituting $\tau + t = z$. Analogously to the previous integral

$$\mathcal{F}^{-1}\left\{-\frac{1}{\mathrm{i}\omega}G(\omega)e^{-\mathrm{i}\omega t}\right\} = -\int_{-\infty}^{x-t} g\,(z)\, dz \tag{15.103e}$$

follows. Finally, the solution in the original space is

$$u(x,t) = \frac{1}{2}f(x+t) + \frac{1}{2}f(x-t) + \int_{x-t}^{x+t} g\,(z)\, dz\,. \tag{15.104}$$

15.4 Z-Transformation

In natural sciences and also in engineering one often has to distinguish between continuous and discrete processes. While continuous processes can be described by differential equations, the discrete processes result mostly in *difference equations*. The solution of differential equations mostly uses Fourier and Laplace transformations, however, to solve difference equations other operator methods have been developed. The best known method is the z-transformation, which is closely related to the Laplace transformation.

15.4.1 Properties of the Z-Transformation

15.4.1.1 Discrete Functions

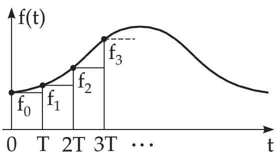


Figure 15.27

If a function $f(t)$ $(0 \le t < \infty)$ is known only at discrete values $t_n = nT$ $(n = 0, 1, 2, \ldots\,;\ T > 0$ is a constant) of the argument, then one writes $f(nT) = f_n$ and forms the sequence $\{f_n\}$. Such a sequence is produced, e.g., in electrotechnics by "scanning" a function $f(t)$ at discrete time periods t_n. Its representation results in a *step function* **(Fig. 15.27)**.
The sequence $\{f_n\}$ and the function $f(nT)$ defined only at discrete points of the argument, which is called a *discrete function*, are equivalent.

15.4.1.2 Definition of the Z-Transformation

1. Original Sequence and Transform

The infinite series

$$F(z) = \sum_{n=0}^{\infty} f_n \left(\frac{1}{z}\right)^n \tag{15.105}$$

is assigned to the sequence $\{f_n\}$. If this series is convergent, then the sequence $\{f_n\}$ is called *z-transformable*, and it is denoted by

$$F(z) = \mathcal{Z}\{f_n\}\,. \tag{15.106}$$

$\{f_n\}$ is called the *original sequence*, $F(z)$ is the *transform*, z denotes a complex variable and $F(z)$ is a complex-valued function.

■ $f_n = 1$ $(n = 0, 1, 2, \ldots)$. The corresponding infinite series is

$$F(z) = \sum_{n=0}^{\infty} \left(\frac{1}{z}\right)^n . \tag{15.107}$$

It represents a geometric series with common ratio $1/z$, which is convergent if $\left|\frac{1}{z}\right| < 1$ and its sum is $F(z) = \frac{z}{z-1}$. It is divergent for $\left|\frac{1}{z}\right| \ge 1$. Therefore, the sequence $\{1\}$ is z-transformable for $\left|\frac{1}{z}\right| < 1$, i.e., for every exterior point of the unit circle $|z| = 1$ in the z plane.

2. Properties

Since the transform $F(z)$ according to (15.105) is a power series of the complex variable $1/z$, the properties of the complex power series (see 14.3.1.3, p. 750) imply the following results:

a) For a z-transformable sequence $\{f_n\}$, there exists a real number R such that the series (15.105) is absolutely convergent for $|z| > 1/R$ and divergent for $|z| < 1/R$. The series is uniformly convergent for $|z| \ge 1/R_0 > 1/R$. R is the radius of convergence of the power series (15.105) of $1/z$. If the series is convergent for every $|z| > 0$, then $R = \infty$. For non z-transformable sequences there is $R = 0$.

b) If $\{f_n\}$ is z-transformable for $|z| > 1/R$, then the corresponding transform $F(z)$ is an analytic function for $|z| > 1/R$ and it is the unique transform of $\{f_n\}$. Conversely, if $F(z)$ is an analytic function

for $|z| > 1/R$ and is regular also at $z = \infty$, then there is a unique original sequence $\{f_n\}$ for $F(z)$. Here, $F(z)$ is called regular at $z = \infty$, if $F(z)$ has a power series expansion in the form (15.105) and $F(\infty) = f_0$.

3. Limit Theorems

Analogously to the limit properties of the Laplace transformation ((15.7b), p. 770), the following limit theorems are valid for the z-transformation:

a) If $F(z) = \mathcal{Z}\{f_n\}$ exists, then

$$f_0 = \lim_{z\to\infty} F(z). \tag{15.108}$$

Here z can tend to infinity along the real axis or along any other path. Since the series

$$z\{F(z) - f_0\} = f_1 + f_2\frac{1}{z} + f_3\frac{1}{z^2} + \cdots\,, \tag{15.109}$$

$$z^2\left\{F(z) - f_0 - f_1\frac{1}{z}\right\} = f_2 + f_3\frac{1}{z} + f_4\frac{1}{z^2} + \cdots\,, \tag{15.110}$$

$$\vdots \qquad\qquad \vdots \qquad\qquad \vdots$$

are obviously z transforms, analogously to (15.108) one gets

$$f_1 = \lim_{z\to\infty} z\{F(z) - f_0\}, \qquad f_2 = \lim_{z\to\infty} z^2\left\{F(z) - f_0 - f_1\frac{1}{z}\right\}, \ldots\,. \tag{15.111}$$

The original sequence $\{f_n\}$ can be determined from its transform $F(z)$ in this way.

b) If $\lim\limits_{n\to\infty} f_n$ exists, then

$$\lim_{n\to\infty} f_n = \lim_{z\to 1+0} (z-1)F(z). \tag{15.112}$$

However the value $\lim\limits_{n\to\infty} f_n$ from (15.112) can be determined only if its existence is guaranteed, since the above statement is not reversible.

■ $f_n = (-1)^n \ (n = 0, 1, 2, \ldots)$. Then $\mathcal{Z}\{f_n\} = \dfrac{z}{z+1}$ and $\lim\limits_{z\to 1+0} (z-1)\dfrac{z}{z+1} = 0$, but $\lim\limits_{n\to\infty} (-1)^n$ does not exist.

15.4.1.3 Rules of Calculations

In applications of the z-transformation it is very important to know how certain operations defined on the original sequences affect the transforms, and conversely. For the sake of simplicity here the notation $F(z) = \mathcal{Z}\{f_n\}$ for $|z| > 1/R$ is used.

1. Translation

Forward and backward translations are distinguished.

1. First Shifting Theorem: $\mathcal{Z}\{f_{n-k}\} = z^{-k}F(z) \qquad (k = 0, 1, 2, \ldots)\,,$ (15.113)

here $f_{n-k} = 0$ is defined for $n - k < 0$.

2. Second Shifting Theorem: $\mathcal{Z}\{f_{n+k}\} = z^k\left[F(z) - \sum\limits_{\nu=0}^{k-1} f_\nu \left(\dfrac{1}{z}\right)^\nu\right] \qquad (k = 1, 2, \ldots)\,.$ (15.114)

2. Summation

For $|z| > \max\left(1, \dfrac{1}{R}\right)$

$$\mathcal{Z}\left\{\sum_{\nu=0}^{n-1} f_\nu\right\} = \frac{1}{z-1}F(z)\,. \tag{15.115}$$

3. Differences
For the *differences*

$$\Delta f_n = f_{n+1} - f_n, \quad \Delta^m f_n = \Delta(\Delta^{m-1} f_n) \qquad (m = 1, 2, \ldots ;\ \Delta^0 f_n = f_n) \tag{15.116}$$

the following equalities hold

$$\begin{aligned} \mathcal{Z}\{\Delta f_n\} &= (z-1)F(z) - z f_0, \\ \mathcal{Z}\{\Delta^2 f_n\} &= (z-1)^2 F(z) - z(z-1) f_0 - z\Delta f_0, \\ \vdots \quad &= \quad \vdots \\ \mathcal{Z}\{\Delta^k f_n\} &= (z-1)^k F(z) - z \sum_{\nu=0}^{k-1} (z-1)^{k-\nu-1} \Delta^\nu f_0. \end{aligned} \tag{15.117}$$

4. Damping
For an arbitrary complex number $\lambda \neq 0$ and $|z| > \dfrac{|\lambda|}{R}$:

$$\mathcal{Z}\{\lambda^n f_n\} = F\left(\frac{z}{\lambda}\right). \tag{15.118}$$

5. Convolution
The *convolution* of two sequences $\{f_n\}$ and $\{g_n\}$ is the operation

$$f_n * g_n = \sum_{\nu=0}^{n} f_\nu\, g_{n-\nu}. \tag{15.119}$$

If the z-transformed functions $\mathcal{Z}\{f_n\} = F(z)$ for $|z| > 1/R_1$ and $\mathcal{Z}\{g_n\} = G(z)$ for $|z| > 1/R_2$ exist, then

$$\mathcal{Z}\{f_n * g_n\} = F(z)G(z) \tag{15.120}$$

for $|z| > \max\left(\dfrac{1}{R_1}, \dfrac{1}{R_2}\right)$. Relation (15.120) is called the *convolution theorem* of the z-transformation. It corresponds to the rules of multiplying two power series.

6. Differentiation of the Transform

$$\mathcal{Z}\{n f_n\} = -z \frac{dF(z)}{dz}. \tag{15.121}$$

Higher-order derivatives of $F(z)$ can be determined by the repeated application of (15.121).

7. Integration of the Transform
Under the assumption $f_0 = 0$,

$$\mathcal{Z}\left\{\frac{f_n}{n}\right\} = \int_z^\infty \frac{F(\xi)}{\xi}\, d\xi. \tag{15.122}$$

15.4.1.4 Relation to the Laplace Transformation

Describing a discrete function $f(t)$ (see 15.4.1.1, p. 794) as a step function, then

$$f(t) = f(nT) = f_n \qquad \text{for} \quad nT \le t < (n+1)T \quad (n = 0, 1, 2, \ldots ;\ T > 0,\ T \ \text{const}) \tag{15.123}$$

holds. Using the Laplace transformation (see 15.2.1.1, **1.**, p. 770) for this piecewise constant function, for $T = 1$ yields:

$$\mathcal{L}\{f(t)\} = F(p) = \sum_{n=0}^{\infty} \int_n^{n+1} f_n e^{-pt}\, dt = \sum_{n=0}^{\infty} f_n \frac{e^{-np} - e^{-(n+1)p}}{p} = \frac{1 - e^{-p}}{p} \sum_{n=0}^{\infty} f_n e^{-np}. \tag{15.124}$$

The infinite series in (15.124) is called the *discrete Laplace transformation* and is denoted by $\mathcal{D}$:

$$\mathcal{D}\{f(t)\} = \mathcal{D}\{f_n\} = \sum_{n=0}^{\infty} f_n e^{-np}. \tag{15.125}$$

After the substitution of $e^p = z$ in (15.125) $\mathcal{D}\{f_n\}$ represents a series with powers of $1/z$, which is a so-called *Laurent series* (see 14.3.4, p. 752). The substitution $e^p = z$ suggested the name of the z transformation. With this substitution from (15.124) one finally gets the following relations between the Laplace and z-transformation in the case of step functions:

$$pF(p) = \left(1-\frac{1}{z}\right)F(z) \quad (15.126\text{a}) \qquad \text{or} \quad p\mathcal{L}\{f(t)\} = \left(1-\frac{1}{z}\right)\mathcal{Z}\{f_n\}. \qquad (15.126\text{b})$$

In this way the relations of z-transforms of step functions (see **Table 21.15**, p. 1128) can be transformed into relations of Laplace transforms of step functions (see **Table 21.13**, p. 1109), and conversely.

15.4.1.5 Inverse of the Z-Transformation

The inverse of the z-transformation is to find the corresponding unique original sequence $\{f_n\}$ from its transform $F(z)$:

$$\mathcal{Z}^{-1}\{F(z)\} = \{f_n\}. \qquad (15.127)$$

There are different possibilities for the inverse transformation.

1. Using Tables

If the function $F(z)$ is not given in tables, then one can try to transform it to a function which is given in **Table 21.15**.

2. Laurent Series of $F(z)$

Using the definition (15.105), p. 794 the inverse transform can be determined directly if a series expansion of $F(z)$ with respect to $1/z$ is known or if it can be determined.

3. Taylor Series of $F\left(\frac{1}{z}\right)$

Since $F\left(\frac{1}{z}\right)$ is a series of increasing powers of z, from (15.105) and using the Taylor formula follows

$$f_n = \frac{1}{n!}\frac{d^n}{dz^n}F\left(\frac{1}{z}\right)\bigg|_{z=0} \quad (n = 0, 1, 2, \ldots). \qquad (15.128)$$

4. Application of Limit Theorems

Using the limits (15.108) and (15.111), p. 795, the original sequence $\{f_n\}$ can be directly determined from its transform $F(z)$.

■ $F(z) = \dfrac{2z}{(z-2)(z-1)^2}$. Using the previous four methods:

1. Partial fraction decomposition (see 1.1.7.3, p. 15) of $F(z)/z$ yields functions which are contained in **Table 21.15**.

$$\frac{F(z)}{z} = \frac{2}{(z-2)(z-1)^2} = \frac{A}{z-2} + \frac{B}{(z-1)^2} + \frac{C}{z-1}. \quad \text{So}$$

$$F(z) = \frac{2z}{z-2} - \frac{2z}{(z-1)^2} - \frac{2z}{z-1} \quad \text{and therefore} \quad f_n = 2(2^n - n - 1) \quad \text{for } n \geq 0.$$

2. By division $F(z)$ gets a series with decreasing powers of z:

$$F(z) = \frac{2z}{z^3 - 4z^2 + 5z - 2} = 2\frac{1}{z^2} + 8\frac{1}{z^3} + 22\frac{1}{z^4} + 52\frac{1}{z^5} + 114\frac{1}{z^6} + \ldots. \qquad (15.129)$$

From this expression one gets $f_0 = f_1 = 0,\ f_2 = 2,\ f_3 = 8,\ f_4 = 22,\ f_5 = 52,\ f_6 = 114,\ \ldots$, but not a closed expression is obtained for the general term f_n.

3. For formulating $F\left(\frac{1}{z}\right)$ and its required derivatives, (see (15.128)) it is advisable to consider the

partial fraction decomposition of $F(z)$

$$\left.\begin{array}{llll} F\left(\frac{1}{z}\right) &= \frac{2}{1-2z} - \frac{2z}{(1-z)^2} - \frac{2}{1-z}, & \text{i.e.} \quad F\left(\frac{1}{z}\right) = 0 & \text{for } z=0, \\ \frac{dF\left(\frac{1}{z}\right)}{dz} &= \frac{4}{(1-2z)^2} - \frac{4z}{(1-z)^3} - \frac{4}{(1-z)^2}, & \text{i.e.} \quad \frac{dF\left(\frac{1}{z}\right)}{dz} = 0 & \text{for } z=0, \\ \frac{d^2F\left(\frac{1}{z}\right)}{dz^2} &= \frac{16}{(1-2z)^3} - \frac{12z}{(1-z)^4} - \frac{12}{(1-z)^3}, & \text{i.e.} \quad \frac{d^2F\left(\frac{1}{z}\right)}{dz^2} = 4 & \text{for } z=0, \\ \frac{d^3F\left(\frac{1}{z}\right)}{dz^3} &= \frac{96}{(1-2z)^4} - \frac{48z}{(1-z)^5} - \frac{48}{(1-z)^4}, & \text{i.e.} \quad \frac{d^3F\left(\frac{1}{z}\right)}{dz^3} = 48 & \text{for } z=0, \\ \vdots & \quad\vdots & \quad\vdots & \quad\vdots \end{array}\right\} \tag{15.130}$$

from which $f_0,\ f_1,\ f_2,\ f_3,\ \ldots$ are easily obtained considering (15.128).

4. Application of the limit theorems (see 15.4.1.2, **3.**, p. 795) gives:

$$f_0 = \lim_{z\to\infty} F(z) = \lim_{z\to\infty} \frac{2z}{z^3 - 4z^2 + 5z - 2} = 0, \tag{15.131a}$$

$$f_1 = \lim_{z\to\infty} z(F(z) - f_0) = \lim_{z\to\infty} \frac{2z^2}{z^3 - 4z^2 + 5z - 2} = 0, \tag{15.131b}$$

$$f_2 = \lim_{z\to\infty} z^2\left(F(z) - f_0 - f_1\frac{1}{z}\right) = \lim_{z\to\infty} \frac{2z^3}{z^3 - 4z^2 + 5z - 2} = 2, \tag{15.131c}$$

$$f_3 = \lim_{z\to\infty} z^3\left(F(z) - f_0 - f_1\frac{1}{z} - f_2\frac{1}{z^2}\right) = \lim_{z\to\infty} z^3\left(\frac{2z}{z^3 - 4z^2 + 5z - 2} - \frac{2}{z^2}\right) = 8, \ \ldots \tag{15.131d}$$

where the *Bernoulli–l'Hospital rule* is applied (see 2.1.4.8, **2.**, p. 56). The original sequence $\{f_n\}$ can be determined successively.

15.4.2 Applications of the Z-Transformation

15.4.2.1 General Solution of Linear Difference Equations

A linear difference equation of order k with constant coefficients has the form

$$a_k y_{n+k} + a_{k-1} y_{n+k-1} + \cdots + a_2 y_{n+2} + a_1 y_{n+1} + a_0 y_n = g_n \quad (n = 0, 1, 2\ldots). \tag{15.132}$$

Here k is a natural number. The coefficients a_i $(i = 0, 1, \ldots, k)$ are given real or complex numbers and they do not depend on n. Here a_0 and a_k are non-zero numbers. The sequence $\{g_n\}$ is given, and the sequence $\{y_n\}$ is to be determined.

To determine a particular solution of (15.132) the values $y_0, y_1, \ldots, y_{k-1}$ have to be previously given. Then the next value y_k can be determined for $n = 0$ from (15.132). Next one gets y_{k+1} for $n = 1$ from $y_1,\ y_2,\ \ldots,\ y_k$ and from (15.132). In this way all values y_n can be calculated recursively. However a general solution can be given for the values y_n with the z-transformation, using the second shifting theorem (15.114) applied for (15.132):

$$a_k z^k \left[Y(z) - y_0 - y_1 z^{-1} - \cdots - y_{k-1} z^{-(k-1)}\right] + \cdots + a_1 z[Y(z) - y_0] + a_0 Y(z) = G(z). \tag{15.133}$$

Here one denotes $Y(z) = \mathcal{Z}\{y_n\}$ and $G(z) = \mathcal{Z}\{g_n\}$. Substituting $a_k z^k + a_{k-1} z^{k-1} + \cdots + a_1 z + a_0 = p(z)$, the solution of the so-called transformed equation (15.133) is

$$Y(z) = \frac{1}{p(z)} G(z) + \frac{1}{p(z)} \sum_{i=0}^{k-1} y_i \sum_{j=i+1}^{k} a_j z^{j-i}. \tag{15.134}$$

As in the case of solving linear differential equations with the Laplace transformation, there is the similar advantage of the z-transformation that initial values are included in the transformed equation, so the solution contains them automatically. The required solution $\{y_n\} = \mathcal{Z}^{-1}\{Y(z)\}$ follows from (15.134) by the inverse transformation discussed in 15.4.1.5, p. 797.

15.4.2.2 Second-Order Difference Equations (Initial Value Problem)

The linear second-order difference equation has the form

$$y_{n+2} + a_1 y_{n+1} + a_0 y_n = g_n \,, \tag{15.135}$$

where y_0 and y_1 are given as initial values. Using the second shifting theorem for (15.135) the transformed equation is

$$z^2\left[Y(z) - y_0 - y_1\frac{1}{z}\right] + a_1 z[Y(z) - y_0] + a_0 Y(z) = G(z). \tag{15.136}$$

Substituting $z^2 + a_1 z + a_0 = p(z)$, the transform is

$$Y(z) = \frac{1}{p(z)}G(z) + y_0\frac{z(z+a_1)}{p(z)} + y_1\frac{z}{p(z)} \,. \tag{15.137}$$

If the roots of the polynomial $p(z)$ are α_1 and α_2, then $\alpha_1 \neq 0$ and $\alpha_2 \neq 0$, otherwise a_0 is zero, and then the difference equation could be reduced to a first-order one. By partial fraction decomposition and applying **Table 21.15** for the z-transformation one gets from

$$\frac{z}{p(z)} = \begin{cases} \dfrac{1}{\alpha_1-\alpha_2}\left(\dfrac{z}{z-\alpha_1} - \dfrac{z}{z-\alpha_2}\right) & \text{for } \alpha_1 \neq \alpha_2, \\ \dfrac{z}{(z-\alpha_1)^2} & \text{for } \alpha_1 = \alpha_2, \end{cases}$$

$$\mathcal{Z}^{-1}\left\{\frac{z}{p(z)}\right\} = \{p_n\} = \begin{cases} \dfrac{\alpha_1^n - \alpha_2^n}{\alpha_1 - \alpha_2} & \text{for } \alpha_1 \neq \alpha_2, \\ n\alpha_1^{n-1} & \text{for } \alpha_1 = \alpha_2. \end{cases} \tag{15.138a}$$

Since $p_0 = 0$, by the second shifting theorem there is

$$\mathcal{Z}^{-1}\left\{\frac{z^2}{p(z)}\right\} = \mathcal{Z}^{-1}\left\{z\,\frac{z}{p(z)}\right\} = \{p_{n+1}\} \tag{15.138b}$$

and by the first shifting theorem

$$\mathcal{Z}^{-1}\left\{\frac{1}{p(z)}\right\} = \mathcal{Z}^{-1}\left\{\frac{1}{z}\,\frac{z}{p(z)}\right\} = \{p_{n-1}\}. \tag{15.138c}$$

Substituting here $p_{-1} = 0$, based on the convolution theorem one gets the original sequence with

$$y_n = \sum_{\nu=0}^{n} p_{\nu-1} g_{n-\nu} + y_0(p_{n+1} + a_1 p_n) + y_1 p_n \,. \tag{15.138d}$$

Since $p_{-1} = p_0 = 0$, this relation and (15.138a) imply that in the case of $\alpha_1 \neq \alpha_2$ it follows

$$y_n = \sum_{\nu=2}^{n} g_{n-\nu}\frac{\alpha_1^{\nu-1} - \alpha_2^{\nu-1}}{\alpha_1 - \alpha_2} + y_0\left(\frac{\alpha_1^{n+1} - \alpha_2^{n+1}}{\alpha_1 - \alpha_2} + a_1\frac{\alpha_1^n - \alpha_2^n}{\alpha_1 - \alpha_2}\right) + y_1\frac{\alpha_1^n - \alpha_2^n}{\alpha_1 - \alpha_2}. \tag{15.138e}$$

This form can be further simplified, since $a_1 = -(\alpha_1 + \alpha_2)$ and $a_0 = \alpha_1\alpha_2$ (see the root theorems of Vieta, 1.6.3.1, 3., p. 44), so

$$y_n = \sum_{\nu=2}^{n} g_{n-\nu}\frac{\alpha_1^{\nu-1} - \alpha_2^{\nu-1}}{\alpha_1 - \alpha_2} - y_0 a_0\frac{\alpha_1^{n-1} - \alpha_2^{n-1}}{\alpha_1 - \alpha_2} + y_1\frac{\alpha_1^n - \alpha_2^n}{\alpha_1 - \alpha_2} \,. \tag{15.138f}$$

In the case of $\alpha_1 = \alpha_2$ similarly

$$y_n = \sum_{\nu=2}^{n} g_{n-\nu}(\nu-1)\alpha_1^{\nu-2} - y_0 a_0 (n-1)\alpha_1^{n-2} + y_1 n \alpha_1^{n-1}. \tag{15.138g}$$

In the case of second-order difference equations the inverse transformation of the transform $Y(z)$ can be performed without partial fraction decomposition using correspondences such as, e.g.,

$$\mathcal{Z}^{-1}\left\{\frac{z}{z^2 - 2az\cosh b + a^2}\right\} = a^{n-1}\frac{\sinh bn}{\sinh n} \tag{15.139}$$

and the second shifting theorem. By substituting $a_1 = -2a\cosh b$, and $a_0 = a^2$ the original sequence of (15.137) becomes:

$$y_n = \frac{1}{\sinh b}\left[\sum_{\nu=2}^{n} g_{n-\nu} a^{\nu-2}\sinh(\nu-1)b - y_0 a^n \sinh(n-1)b + y_1 a^{n-1}\sinh nb\right]. \tag{15.140}$$

This formula is useful in numerical computations especially if a_0 and a_1 are complex numbers.

Remark: Notice that the hyperbolic functions are also defined for complex variables.

15.4.2.3 Second-Order Difference Equations (Boundary Value Problem)

It often happens in applications that the values y_n of a difference equation are needed only for a finite number of indices $0 \le n \le N$. In the case of a second-order difference equation (15.135) both *boundary values* y_0 and y_N are usually given. To solve this boundary value problem one starts with the solution (15.138f) of the corresponding initial value problem, where instead of the unknown value y_1 it is to introduce y_N. Substituting $n = N$ into (15.138f), y_1 can be obtained which depends on y_0 and y_N:

$$y_1 = \frac{1}{\alpha_1^N - \alpha_2^N}\left[y_0 a_0(\alpha_1^{N-1} - \alpha_2^{N-1}) + y_N(\alpha_1 - \alpha_2) - \sum_{\nu=2}^{N}(\alpha_1^{\nu-1} - \alpha_2^{\nu-1})g_{N-\nu}\right]. \tag{15.141}$$

Substituting this value into (15.138f)

$$\begin{aligned} y_n = {} & \frac{1}{\alpha_1 - \alpha_2}\sum_{\nu=2}^{n}(\alpha_1^{\nu-1} - \alpha_2^{\nu-1})g_{n-\nu} - \frac{1}{\alpha_1 - \alpha_2}\frac{\alpha_1^n - \alpha_2^n}{\alpha_1^N - \alpha_2^N}\sum_{\nu=2}^{N}(\alpha_1^{\nu-1} - \alpha_2^{\nu-1})g_{N-\nu} \\ & + \frac{1}{\alpha_1^N - \alpha_2^N}[\,y_0(\alpha_1^N\alpha_2^n - \alpha_1^n\alpha_2^N) + y_N(\alpha_1^n - \alpha_2^n)]. \end{aligned} \tag{15.142}$$

The solution (15.142) makes sense only if $\alpha_1^N - \alpha_2^N \neq 0$ holds. Otherwise, the boundary value problem has no general solution, but analogously to the boundary value problems of differential equations eigenvalues and eigenfunctions emerge.

15.5 Wavelet Transformation

15.5.1 Signals

If a physical object emits an effect which spreads out and can be described mathematically, e.g., by a function or a number sequence, then it is called a *signal*.

Signal analysis means to characterize a signal by a quantity that is typical for the signal. This means mathematically: The function or the number sequence, which describes the signal, will be mapped into another function or number sequence, from which the typical properties of the signal can be clearly seen. For such mappings, of course, some informations can also be lost.

The reverse operation of signal analysis, i.e., the reconstruction of the original signal, is called *signal synthesis*.

The connection between signal analysis and signal synthesis can be well represented by an example of Fourier transformation: A signal $f(t)$ (t denotes time) is characterized by the frequency ω. Then, formula (15.143a) describes the signal analysis, and formula (15.143b) describes the signal synthesis:

$$F(\omega) = \int_{-\infty}^{\infty} e^{-i\omega t} f(t)\,dt \quad (15.143a) \quad \text{and} \quad f(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{i\omega t} F(\omega)\,d\omega. \quad (15.143b)$$

15.5.2 Wavelets

The Fourier transformation has no localization property, i.e., if a signal changes at one position, then the transform changes everywhere without the possibility that the position of the change could be recognized "at a glance". The basis of this fact is that the Fourier transformation decomposes a signal into *plane waves*. These are described by trigonometric functions, which oscillate with the same period for arbitrary long time. However, for wavelet transformations there is an almost freely chosen function ψ, the *wavelet* (small localized wave), that is shifted and compressed for analysing a signal.

Examples are the Haar wavelet **(Fig. 15.28a)** and the Mexican hat **(Fig. 15.28b)**.

■ **A Haar wavelet:**

$$\psi = \begin{cases} 1 & \text{if} \quad 0 \le x < \frac{1}{2}, \\ -1 & \text{if} \quad \frac{1}{2} \le x \le 1, \\ 0 & \text{otherwise.} \end{cases} \quad (15.144)$$

■ **B Mexican hat:**

$$\psi(x) = -\frac{d^2}{dx^2} e^{-x^2/2} \quad (15.145)$$

$$= (1 - x^2)e^{-x^2/2}. \quad (15.146)$$

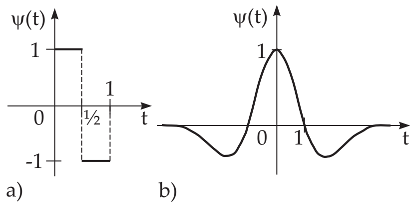


Figure 15.28

Generally, it holds that every function ψ comes into consideration as a wavelet if it is quadratically integrable and its Fourier transform $\Psi(\omega)$ according to (15.143a) results in a positive finite integral

$$\int_{-\infty}^{\infty} \frac{|\Psi(\omega)|}{|\omega|}\,d\omega. \quad (15.147)$$

Concerning wavelets, the following properties and definitions are essential:

1. For the mean value of the wavelet:

$$\int_{-\infty}^{\infty} \psi(t)\,dt = 0. \quad (15.148)$$

2. The following integral is called the k-th moment of a wavelet ψ:

$$\mu_k = \int_{-\infty}^{\infty} t^k \psi(t)\,dt. \quad (15.149)$$

The smallest positive integer n such that $\mu_n \ne 0$, is called the *order* of the wavelet ψ.

■ For the Haar wavelet (15.144), $n = 1$, and for the Mexican hat (15.146), $n = 2$.

3. When $\mu_k = 0$ for every k, ψ has infinite order. Wavelets with bounded support always have finite order.

4. A wavelet of order n is orthogonal to every polynomial of degree $\le n - 1$.

15.5.3 Wavelet Transformation

For a wavelet $\psi(t)$ a family of curves can be formed with parameter a:

$$\psi_a(t) = \frac{1}{\sqrt{|a|}}\psi\left(\frac{t}{a}\right) \quad (a \ne 0). \quad (15.150)$$

In the case of $|a| > 0$ the initial function $\psi(t)$ is compressed. In the case of $a < 0$ there is an additional reflection. The factor $1/\sqrt{|a|}$ is a scaling factor.

The functions $\psi_a(t)$ can also be shifted by a second parameter b. Then a two-parameter family of curves arises:

$$\psi_{a,b} = \frac{1}{\sqrt{|a|}}\psi\left(\frac{t-b}{a}\right) \qquad (a, b \text{ real}; \quad a \neq 0). \tag{15.151}$$

The real shifting parameter b characterizes the first moment, while parameter a gives the deviation of the function $\psi_{a,b}(t)$. The function $\psi_{a,b}(t)$ is called a *basis function* in connection to the *wavelet transformation*.

The wavelet transformation of a function $f(t)$ is defined as:

$$\mathcal{L}_\psi f(a,b) = c\int\limits_{-\infty}^{\infty} f(t)\psi_{a,b}(t)\,dt = \frac{c}{\sqrt{|a|}}\int\limits_{-\infty}^{\infty} f(t)\psi\left(\frac{t-b}{a}\right)dt. \tag{15.152a}$$

For the inverse transformation:

$$f(t) = c\int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} \mathcal{L}_\psi f(t)\psi_{a,b}(t)\frac{1}{a^2}\,da\,db. \tag{15.152b}$$

Here c is a constant dependent on the special wavelet ψ.

■ Using the Haar wavelets (15.146) gives

$$\psi\left(\frac{t-b}{a}\right) = \begin{cases} 1 & \text{if } \; b \le t < b + a/2, \\ -1 & \text{if } \; b + a/2 \le t < b + a, \\ 0 & \quad\;\; \text{otherwise} \end{cases}$$

and therefore

$$\begin{aligned} \mathcal{L}_\psi f(a,b) &= \frac{1}{\sqrt{|a|}}\left(\int_b^{b+a/2} f(t)\,dt - \int_{b+a/2}^{b+a} f(t)\,dt\right) \\ &= \frac{\sqrt{|a|}}{2}\left(\frac{2}{a}\int_b^{b+a/2} f(t)\,dt - \frac{2}{a}\int_{b+a/2}^{b+a} f(t)\,dt\right). \end{aligned} \tag{15.153}$$

The value $\mathcal{L}_\psi f(a,b)$ given in (15.153) represents the difference of the mean values of a function $f(t)$ over two neighboring intervals of length $\frac{|a|}{2}$, connected at the point b.

Remarks:

1. The *dyadic wavelet transformation* has an important role in applications. As basis functions are used the functions

$$\psi_{i,j}(t) = \frac{1}{\sqrt{2^i}}\psi\left(\frac{t-2^i j}{2^i}\right), \tag{15.154}$$

i.e., different basis functions can be generated from one wavelet $\psi(t)$ by doubling or halving the width and shifting by an integer multiple of the width.

2. A wavelet $\psi(t)$ is called an *orthogonal wavelet*, if the basis functions given in (15.154) form an orthogonal system.

3. The Daubechies wavelets have especially good numerical properties. They are orthogonal wavelets with compact support, i.e., they are different from zero only on a bounded subset of the time scale.

They do not have a closed form representation (see [15.9]).

15.5.4 Discrete Wavelet Transformation

15.5.4.1 Fast Wavelet Transformation

The integral representation (15.152b) is very redundant, and so the double integral can be replaced by a double sum without loss of information. Considering this idea at the concrete application of the wavelet transformation one needs

1. an efficient algorithm of the transformation, which leads to the concept of *multi-scale analysis*, and

2. an efficient algorithm of the inverse transformation, i.e., an efficient way to reconstruct signals from their wavelet transformations, which leads to the concept of *frames*.

For more details about these concepts see [15.9], [15.1].

Remark: The great success of wavelets in many different applications, such as
- calculation of physical quantities from measured sequences
- pattern and voice recognition
- data compression in news transmission

is based on "fast algorithms". Analogously to the **FFT** (**F**ast **F**ourier **T**ransformation, see 19.6.4.2, p. 993) one talks here about **FWT** (**F**ast **W**avelet **T**ransformation).

15.5.4.2 Discrete Haar Wavelet Transformation

An example of a discrete wavelet transformation is the Haar wavelet transformation: The values f_i $(i = 1, 2, \ldots, N)$ are given from a signal. The detailed values d_i $(i = 1, 2, \ldots, N/2)$ are calculated as:

$$s_i = \frac{1}{\sqrt{2}}(f_{2i-1} + f_{2i}), \quad d_i = \frac{1}{\sqrt{2}}(f_{2i-1} - f_{2i}). \tag{15.155}$$

The values d_i are to be stored while the rule (15.155) is applied to the values s_i, i.e., in (15.155) the values f_i are replaced by the values s_i. This procedure is continued, sequentially so that finally from

$$s_i^{(n+1)} = \frac{1}{\sqrt{2}}\left(s_{2i-1}^{(n)} + s_{2i}^{(n)}\right), \quad d_i^{(n+1)} = \frac{1}{\sqrt{2}}\left(s_{2i-1}^{(n)} - s_{2i}^{(n)}\right) \tag{15.156}$$

a sequence of detailed vectors is formed with components $d_i^{(n)}$. Every detailed vector contains information about the properties of the signals.

Remark: For large values of N the discrete wavelet transformation converges to the integral wavelet transformation (15.152a).

15.5.5 Gabor Transformation

Time-frequency analysis is the characterization of a signal with respect to the contained frequencies and time periods when these frequencies appear. Therefore, the signal is divided into time segments (windows) and a Fourier transform is used. It is called a **W**indowed **F**ourier **T**ransformation (**WFT**).

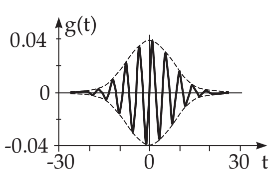


Figure 15.29

The window function should be chosen so that a signal is considered only in the window. Gabor applied the window function

$$g(t) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{t^2}{2\sigma^2}} \tag{15.157}$$

(Fig. 15.29). This choice can be explained as $g(t)$, with the "total unit mass", is concentrated at the point $t = 0$ and the width of the window can be considered as a constant (about 2σ).

The *Gabor transformation* of a function $f(t)$ then has the form

$$\mathcal{G}f(\omega, s) = \int_{-\infty}^{\infty} f(t)g(t-s)e^{-\mathrm{i}\omega t}\, dt. \qquad (15.158)$$

This determines, with which complex amplitude the dominant wave (fundamental harmonic) $e^{\mathrm{i}\omega t}$ occurs during the time interval $[s-\sigma, s+\sigma]$ in f, i.e., if the frequency ω occurs in this interval, then it has the amplitude $|\mathcal{G}f(\omega, s)|$.

15.6 Walsh Functions

15.6.1 Step Functions

Orthogonal systems of functions have an important role in the approximation theory of functions. For instance, special polynomials or trigonometric functions are used since they are smooth, i.e., they are differentiable sufficiently many times in the considered interval. However, there are problems, e.g., the transition of points of a rough picture, when smooth functions are not suitable for the mathematical description, but *step functions*, piecewise constant functions are more appropriate. Walsh functions are very simple step functions. They take only two function values +1 and −1. These two function values correspond to two states, so the Walsh functions can be implemented by computers very easily.

15.6.2 Walsh Systems

Analogously to trigonometric functions also periodic step functions can be considered. The interval $I = [0, 1)$ is used as a period interval and it is divided into 2^n equally long subintervals. Suppose S_n is the set of periodic step functions with period 1 over such an interval. The different step functions belonging to S_n can be considered as vectors of a finite dimensional vector space, since every function $g \in S_n$ is defined by its values $g_0, g_1, g_2, \ldots, g_{2^n-1}$ in the subintervals and it can be considered as a vector:

$$\underline{\mathbf{g}}^{\mathrm{T}} = (g_0, g_1, g_2, \ldots, g_{2^n-1}). \qquad (15.159)$$

The Walsh functions belonging to S_n form an orthogonal basis with respect to a suitable scalar product in this space. The basis vectors can be enumerated in many different ways, so one can get many different Walsh systems, which actually contain the same functions. There are three of them which should be mentioned: Walsh–Kronecker functions, Walsh–Kaczmarz functions and Walsh–Paley functions.

The *Walsh transformation* is constructed analogously to the Fourier transformation, where the role of the trigonometric functions is taken by the Walsh functions. One gets, e.g., Walsh series, Walsh polynomials, Walsh sine and Walsh cosine transformations, Walsh integral, and analogously to the fast Fourier transformation there is a Fast Walsh Transformation. For an introduction in the theory and applications of Walsh functions see [15.6].

16 Probability Theory and Mathematical Statistics

When experiments or observations are made, various outcomes are possible even under the same conditions. Probability theory and statistics deal with regularity of random outcomes of certain results with respect to given experiments or observations. (In probability theory and statistics, observations are also called experiments, since they have certain outcomes.) It is supposed, at least theoretically, that these experiments can be repeated arbitrarily many times under the same circumstances. Its application have these disciplines of mathematics in the statistic assessment of mass phenomena. The mathematical handling of *random phenomena* is also summarized in the notion *stochastics*.

16.1 Combinatorics

From the elements of a given set often new sets, systems or sequences are composed. Depending on the way how to do it, one gets the notions *permutation* (ordering), *combination* (selection), and *partial permutation* or *arrangement*. The notion arrangement combines ordering and selection. The basic problem of combinatorics is to determine how many different choices or arrangements are possible with the given elements.

16.1.1 Permutations

1. Definition

A *permutation* of n elements is an ordering of the n elements.

2. Number of Permutations without Repetition

The number of different permutations of n different elements is

$$P_n = n!\,. \tag{16.1}$$

■ In a classroom 16 students are seated on 16 places. There are 16! different possibilities for seating.

3. Number of Permutations with Repetitions

The number ${P_n}^{(k)}$ of different permutations of n elements containing k identical elements $(k \le n)$ is

$${P_n}^{(k)} = \frac{n!}{k!}\,. \tag{16.2}$$

■ In a classroom 16 school-bags of 16 students are placed on 16 chairs. Four of them are identical. There are 16!/4! different placements of the school-bags.

4. Generalization

The number ${P_n}^{(k_1,k_2,\ldots,k_m)}$ of different permutations of n elements containing m different types of elements with multiplicities $k_1, k_2, \ldots, k_m$ respectively $(k_1 + k_2 + \ldots + k_m = n)$ is

$${P_n}^{(k_1,k_2,\ldots,k_m)} = \frac{n!}{k_1!k_2!\ldots k_m!}\,. \tag{16.3}$$

■ Suppose one composes five-digit numbers from the digits 4, 4, 5, 5, 5. One can compose ${P_5}^{(2,3)} = \frac{5!}{2!3!} = 10$ different numbers.

16.1.2 Combinations

1. Definition

A *combination* is a choice of k elements from n different elements not considering the order of them. This is called a combination of k-th order and one distinguishes between combinations with and without repetition.

I.N. Bronshtein et al., *Handbook of Mathematics*,
DOI 10.1007/978-3-662-46221-8_16

2. Number of Combinations without Repetition

The number ${C_n}^{(k)}$ of different possibilities to choose k elements from n different elements not considering the order is

$$ {C_n}^{(k)} = \binom{n}{k} \quad \text{with} \quad 0 \le k \le n \quad \text{(see binomial coefficient in 1.1.6.4, \textbf{3.}, p. 13)}, \tag{16.4} $$

if any element is chosen at most once. This is called a combination without repetition.

■ There are $\binom{30}{4} = 27405$ possibilities to choose an electoral board of four persons from 30 participants.

3. Number of Combinations with Repetition

The number of possibilities to choose k elements from n different ones, repeating each element arbitrarily times and not considering the order is

$$ {C_n}^{(k)} = \binom{n+k-1}{k}. \tag{16.5} $$

In other words, the number of different selections of k elements chosen from n different elements is considered, where each of n can be chosen more then once.

■ Rolling k dice, ${C_6}^{(k)} = \binom{k+6-1}{k}$ different results are possible. Consequently, for two dice there are ${C_6}^{(2)} = \binom{7}{2} = 21$ different results.

16.1.3 Arrangements

1. Definition

An *arrangement* is an ordering of k elements selected from n different ones, i.e., arrangements are combinations considering the order.

2. Number of Arrangements without Repetition

The number ${V_n}^{(k)}$ of different orderings of k different elements selected from n different ones is

$$ {V_n}^{(k)} = k!\binom{n}{k} = n(n-1)(n-2)\dots(n-k+1) \quad (0 \le k \le n). \tag{16.6} $$

■ How many different ways are there to choose a chairman, his deputy, and a first and a second assistant for them from 30 participants at an election meeting? The answer is $\binom{30}{4}4! = 657720$.

3. Number of Arrangements with Repetition

An ordering of k elements selected from n different ones, where any of the elements can be selected arbitrarily many times, is called an arrangement with repetition. Their number is

$$ {V_n}^{(k)} = n^k. \tag{16.7} $$

■ **A:** In a soccer-toto with 12 games there are 3^{12} different outcomes.

■ **B:** With the digital unit called a byte which contains 8 bits can be represented $2^8 = 256$ different symbols (see for example the well-known ASCII table).

16.1.4 Collection of the Formulas of Combinatorics (see Table 16.1)

Table 16.1 Collection of the formulas of combinatorics

Type of choice or selection of k from n elements	Number of possibilities without repetition ($k \le n$)	with repetition ($k \le n$)
Permutations	$P_n = n!\ (n = k)$	$P_n^{(k)} = \dfrac{n!}{k!}$
Combinations	$C_n{}^{(k)} = \binom{n}{k}$	$C_n{}^{(k)} = \binom{n+k-1}{k}$
Arrangements	$V_n{}^{(k)} = k!\binom{n}{k}$	$V_n{}^{(k)} = n^k$

16.2 Probability Theory

16.2.1 Event, Frequency and Probability

16.2.1.1 Events

1. Different Types of Events

All the possible outcomes of an experiment are called *events* in probability theory, and they form the *fundamental probability set* **A**.

There are to be distinguished the *certain event*, the *impossible event* and *random events*.

The certain event occurs every time when the experiment is performed, the impossible event never occurs; a random event sometimes occurs, sometimes does not. All possible outcomes of the experiment excluding each other are called *elementary events* (see also **Table 16.2**). Here the events of the fundamental probability set **A** are denoted by $A, B, C, \ldots$, the certain event by I, the impossible event by O. Some operations and relations between the events are denoted as given in **Table 16.2**.

2. Properties of the Operations

The fundamental probability set forms a Boolean algebra with complement, addition, and multiplication defined in **Table 16.2**, and it is called the *field of events*. The following rules are valid:

1. a) $A + B = B + A,$ (16.8) **1. b)** $AB = BA.$ (16.9)

2. a) $A + A = A,$ (16.10) **2. b)** $AA = A.$ (16.11)

3. a) $A + (B + C) = (A + B) + C,$ (16.12) **3. b)** $A(BC) = (AB)C.$ (16.13)

4. a) $A + \overline{A} = I,$ (16.14) **4. b)** $A\overline{A} = O.$ (16.15)

5. a) $A(B + C) = AB + AC,$ (16.16) **5. b)** $A + BC = (A + B)(A + C).$ (16.17)

6. a) $\overline{A + B} = \overline{A}\,\overline{B},$ (16.18) **6. b)** $\overline{AB} = \overline{A} + \overline{B}.$ (16.19)

7. a) $B - A = B\overline{A},$ (16.20) **7. b)** $\overline{A} = I - A.$ (16.21)

8. a) $A(B - C) = AB - AC,$ (16.22) **8. b)** $AB - C = (A - C)(B - C).$ (16.23)

9. a) $O \subseteq A,$ (16.24) **9. b)** $A \subseteq I.$ (16.25)

10. From $A \subseteq B$ follows **a)** $A = AB$ (16.26) and **b)** $B = A + B\overline{A}$ and conversely. (16.27)

11. Complete System of Events: A system of events A_α ($\alpha \in \theta$, θ is a finite or infinite set of indices) is called a *complete system of events* if the following is valid:

11. a) $A_\alpha A_\beta = O$ for $\alpha \neq \beta$ (16.28) and **11. b)** $\sum\limits_{\alpha\in\theta} A_\alpha = I.$ (16.29)

Table 16.2 Relations between events

	Name	**Notation**	**Definition**
1.	Complementary event of A:	$\overline{A}$	$\overline{A}$ occurs exactly if A does not.
2.	Sum of events A and B:	$A + B$	$A+B$ is the event which occurs if A or B or both occur.
3.	Product of the events A and B:	AB	AB is the event which occurs exactly if both A and B occur.
4.	Difference of the events A and B:	$A - B$	$A - B$ occurs exactly if A occurs and B does not.
5.	Event as a consequence of the other:	$A \subseteq B$	$A \subseteq B$ means that from the occurrence of A follows the occurrence of B.
6.	Elementary or simple event:	E	From $E = A + B$ it follows that $E = A$ or $E = B$.
7.	Compound event:		Event, which is not elementary.
8.	Disjoint or exclusive events A and B:	$AB = O$	The events A and B cannot occur at the same time.

■ **A:** Tossing two coins: Elementary events for the separate tossing: See the table on the right.

	Head	Tail
1. Coin	A_{11}	A_{12}
2. Coin	A_{21}	A_{22}

1. Elementary event for tossing both coins, e.g.: First coin shows head, second shows tail: $A_{11}A_{22}$.

Compound event for tossing both coins: First coin shows head: $A_{11} = A_{11}A_{21} + A_{11}A_{22}$

2. Compound event for tossing one coin, e.g., the first one: First coin shows head or tail: $A_{11} + A_{12} = I$. Head and tail on the same coin are disjoint events: $A_{11}A_{12} = O$.

■ **B:** Lifetime of light-bulbs.
Defining the elementary events A_n: the lifetime t satisfies the inequalities $(n-1)\Delta t < t \leq n\Delta t$ ($n = 1, 2, \ldots$, and $\Delta t > 0$, arbitrary unit of time).
Compound event A: The lifetime is at most $n\Delta t$, i.e., $A = \sum\limits_{\nu=1}^{n} A_\nu$.

16.2.1.2 Frequencies and Probabilities

1. Frequencies

Let A be an event belonging to the field of events $\mathbf{A}$ of an experiment. If event A occurred n_A times repeating the experiment n times, then n_A is called the *frequency*, and $n_A/n = h_A$ is called the *relative frequency* of the event A. The relative frequency satisfies certain properties which can be used to built up an axiomatic definition of the notion of the probability $P(A)$ of event A in the field of events $\mathbf{A}$.

2. Definition of the Probability

A real function P defined on the field of events is called a *probability* if it satisfies the following properties:

1. For every event $A \in \mathbf{A}$

$$0 \le P(A) \le 1, \quad \text{and} \quad 0 \le h_A \le 1. \tag{16.30}$$

2. For the impossible event O and the certain event I

$$P(O) = 0, \ P(I) = 1, \quad \text{and} \quad h_O = 0, \ h_I = 1. \tag{16.31}$$

3. If the events $A_i \in \mathbf{A}$ $(i = 1, 2, \ldots)$ are finite or countably many mutually exclusive events ($A_i A_k = O$ for $i \neq k$), then

$$P(A_1 + A_2 + \ldots) = P(A_1) + P(A_2) + \ldots, \quad \text{and} \quad h_{A_1+A_2+\ldots} = h_{A_1} + h_{A_2} + \ldots. \tag{16.32}$$

Remark: The axiomatization of the probability theory, which is based on the assumption of three conditions of the preceding kind, succeeded in 1933 A.N. Kolmogoroff (see [16.11]).

3. Rules for Probabilities

1. $B \subseteq A$ yields $P(B) \le P(A)$. (16.33)

2. $P(A) + P(\overline{A}) = 1$. (16.34)

3.a) For n mutually exclusive events A_i $(i = 1, \ldots, n;\, A_i A_k = O, i \neq k)$, holds

$$P(A_1 + A_2 + \ldots A_n) = P(A_1) + P(A_2) + \ldots + P(A_n). \tag{16.35a}$$

3.b) In particular for $n = 2$ holds

$$P(A + B) = P(A) + P(B). \tag{16.35b}$$

4. a) For arbitrary events A_i $(i = 1, \ldots, n)$, holds

$$\begin{aligned} P(A_1 + \cdots + A_n) = {} & P(A_1) + \cdots + P(A_n) - P(A_1A_2) - \cdots - P(A_1A_n) \\ & - P(A_2A_3) - \cdots - P(A_2A_n) - \cdots - P(A_{n-1}A_n) \\ & + P(A_1A_2A_3) + \cdots + P(A_1A_2A_n) + \cdots + P(A_{n-2}A_{n-1}A_n) - \\ & \vdots \\ & + (-1)^{n-1} P(A_1A_2 \ldots A_n). \end{aligned} \tag{16.36a}$$

4.b) In particular for $n = 2$ holds $P(A_1 + A_2) = P(A_1) + P(A_2) - P(A_1A_2)$. (16.36b)

5. Equally likely events: If every event A_i $(i = 1, 2, \ldots, n)$ of a finite complete system of events occurs with the same probability, then

$$P(A_i) = \frac{1}{n}. \tag{16.37}$$

If A is a sum of m $(m \le n)$ events A_i $(i = 1, 2, \ldots, n)$ with the same probability of a complete system, then

$$P(A) = \frac{m}{n}. \tag{16.38}$$

4. Examples of Probabilities

■ **A:** The probability $P(A)$ to get a 2 rolling a fair die is: $P(A) = \dfrac{1}{6}$.

■ **B:** What is the probability of guessing four numbers for the lotto "6 from 49", i.e., 6 numbers are to be chosen from the numbers $1, 2, \ldots, 49$.

If 6 numbers are drawn, then there are $\binom{6}{4}$ possibilities to choose 4. On the other hand there are $\binom{49-6}{6-4} = \binom{43}{2}$ possibilities for the false numbers. Altogether, there are $\binom{49}{6}$ different possibilities to draw 6 numbers. Therefore, the probability $P(A_4)$ is:

$$P(A_4) = \frac{\binom{6}{4}\binom{43}{2}}{\binom{49}{6}} = \frac{645}{665896} = 0.0968\,\%.$$

Similarly, the probability $P(A_6)$ to get a direct hit is:

$$P(A_6) = \frac{1}{\binom{49}{6}} = 0.715 \cdot 10^{-7} = 7.15 \cdot 10^{-6}\,\%.$$

■ **C:** What is the probability $P(A)$ that at least two persons have birthdays on the same day among k persons? (The years of birth must not be identical, and one supposes that every day has the same probability of being a birthday.)

It is easier to consider the complementary event $\overline{A}$: All the k persons have different birthdays. One gets:

$$P(\overline{A}) = \frac{365}{365} \cdot \frac{365-1}{365} \cdot \frac{365-2}{365} \cdot \ldots \cdot \frac{365-k+1}{365}.$$

From this it follows that

$$P(A) = 1 - P(\overline{A}) = 1 - \frac{365 \cdot 364 \cdot 363 \cdot \ldots \cdot (365-k+1)}{365^k}.$$

Some numerical results:

k	10	20	23	30	60
P(A)	0.117	0.411	0.507	0.706	0.994

It is to see that the probability that among 23 and more persons at least two have the same birthday is greater than 50 %.

16.2.1.3 Conditional Probability, Bayes Theorem

1. Conditional Probability

The probability of an event B, when it is known that some event A has already occurred, is called a *conditional probability* and it is denoted by $P(B|A)$ or $P_A(B)$ (read: The probability that B occurs given that A has occurred). It is defined by

$$P(B|A) = \frac{P(AB)}{P(A)}, \quad P(A) \neq 0. \tag{16.39}$$

The conditional probability satisfies the following properties:

a) If $P(A) \neq 0$ and $P(B) \neq 0$ holds, then

$$\frac{P(B|A)}{P(B)} = \frac{P(A|B)}{P(A)}. \tag{16.40a}$$

b) If $P(A_1A_2A_3 \ldots A_n) \neq 0$ holds, then

$$P(A_1A_2 \ldots A_n) = P(A_1)P(A_2|A_1) \ldots P(A_n|A_1A_2 \ldots A_{n-1}). \tag{16.40b}$$

2. Independent Events

The events A and B are *independent events* if

$$P(A|B) = P(A) \quad \text{and} \quad P(B|A) = P(B) \tag{16.41a}$$

holds. In this case, it is

$$P(AB) = P(A)P(B). \tag{16.41b}$$

3. Events in a Complete System of Events

If $\mathbf{A}$ is a field of events and the events $B_i \in \mathbf{A}$ with $P(B_i) > 0$ $(i = 1, 2, \ldots)$ form a complete system of events, then for an arbitrary event $A \in \mathbf{A}$ the following formulas are valid:

a) Total Probability Theorem

$$P(A) = \sum_i P(A|B_i)P(B_i). \tag{16.42}$$

b) Bayes Theorem with $P(A) > 0$

$$P(B_k|A) = \frac{P(A|B_k)P(B_k)}{\sum\limits_i P(A|B_i)P(B_i)} . \tag{16.43}$$

■ Three machines produce the same type of product in a factory. The first one gives 20 % of the total production, the second one gives 30 % and the third one 50 %. It is known from past experience that 5 %, 4 %, and 2 % of the product made by each machine, respectively, are defective. Two types of questions often arise:

a) What is the probability that an article selected randomly from the total production is defective?

b) If the randomly selected article is defective, what is the probability that it was made, e.g., by the first machine?

Using the following notation gives:

• A_i denotes the event that the randomly selected article is made by the i-th machine ($i = 1, 2, 3$) with $P(A_1) = 0.2$, $P(A_2) = 0.3$, $P(A_3) = 0.5$. The events A_i form a complete system of events:

• $A_iA_j = 0,\ A_1 + A_2 + A_3 = I$.

• A denotes the event that the chosen article is defected.

• $P(A|A_1) = 0.05$ gives the probability that an article produced by the first machine is defective; analogously $P(A|A_2) = 0.04$ and $P(A|A_3) = 0.02$ hold.

Now, the answers to the questions are:

a) $P(A) = P(A_1)P(A|A_1) + P(A_2)P(A|A_2) + P(A_3)P(A|A_3)$
$\quad = 0.2 \cdot 0.05 + 0.3 \cdot 0.04 + 0.5 \cdot 0.02 = 0.032.$

b) $P(A_1|A) = P(A_1)\dfrac{P(A|A_1)}{P(A)} = 0.2\dfrac{0.05}{0.032} = 0.31.$

16.2.2 Random Variables, Distribution Functions

To apply the methods of analysis in probability theory, the notions of variable and function are necessary.

16.2.2.1 Random Variable

A set of elementary events is to be described by a *random variable* X. The random variable X can be considered as a quantity which takes its values x randomly from a subset R of the real numbers.
If R contains finitely or countably many different values, then X is called a *discrete random variable*. In the case of a *continuous random variable*, R can be the whole real axis or it may contain subintervals. For the precise definition see 16.2.2.2, **2.**, p. 812. There are also *mixed random variables*.

■ **A:** Assigning the values $1, 2, 3, 4$ to the elementary events $A_{11}, A_{12}, A_{21}, A_{22}$, respectively, in example **A**, p. 808, then a discrete random variable X is defined.

■ **B:** The lifetime T of a randomly selected light-bulb is a continuous random variable. The elementary event $T = t$ occurs if the lifetime T is equal to t.

16.2.2.2 Distribution Function

1. Distribution Function and its Properties

The distribution of a random variable X can be given by its distribution function

$$F(x) = P(X \le x) \quad \text{for} \quad -\infty \le x \le \infty. \tag{16.44}$$

It determines the probability that the random variable X takes a value between $-\infty$ and x. Its domain is the whole real axis. The distribution function has the following properties:

1. $F(-\infty) = 0,\quad F(+\infty) = 1.$

2. $F(x)$ is a non-decreasing function of x.

3. $F(x)$ is continuous on the right.

Remarks:

1. From the definition it follows that $P(X = a) = F(a) - \lim\limits_{x \to a-0} F(x)$.

2. In the literature, also the definition $F(x) = P(X < x)$ is often used. In this case $P(X = a) = \lim\limits_{x \to a+0} F(x) - F(a)$.

2. Distribution Function of Discrete and Continuous Random Variables

a) Discrete Random Variable: A discrete random variable X, which takes the values x_i $(i = 1, 2, \ldots)$ with probabilities $P(X = x_i) = p_i$ $(i = 1, 2, \ldots)$, has the distribution function

$$F(x) = \sum_{x_i \le x} p_i. \tag{16.45}$$

b) Continuous Random Variable: A random variable is called continuous if there exists a non-negative function $f(x)$ such that the probability $P(X \in S)$ can be expressed as $P(X \in S) = \int_S f(x)dx$ for any domain S such that it is possible to consider an integral over it. This function is the so-called *density function*. A continuous random variable takes any given value x_i with 0 probability, so one rather considers the probability that X takes its value from a finite interval $[a, b]$:

$$P(a \le X \le b) = \int_a^b f(t)\,dt. \tag{16.46}$$

A continuous random variable has an everywhere *continuous distribution function*:

$$F(x) = P(X \le x) = \int_{-\infty}^{x} f(t)\,dt. \tag{16.47}$$

$F'(x) = f(x)$ holds at the points where $f(x)$ is continuous.

Remark: When there is no confusion about the upper integration limit, often the integration variable is denoted by x instead of t.

3. Area Interpretation of the Probability, Quantile

By introducing the distribution function and density function in (16.47), the probability $P(X \le x) = F(x)$ can be represented as an area between the density function $f(t)$ and the x-axis on the interval $-\infty < t \le x$ **(Fig. 16.1a)**.

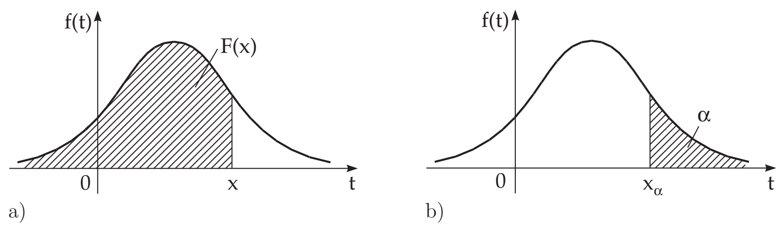


Figure 16.1

Often there is given (frequently in %) a probability value α. If

$$P(X > x) = \alpha \tag{16.48}$$

holds, the corresponding value of the abscissa $x = x_\alpha$ is called the *quantile* or the *fractile of order* α **(Fig. 16.1b)**. This means the area under the density function $f(t)$ to the right of x_α is equal to α.

Remark: In the literature, the area to the left of x_α is also used for the definition of quantile.

In mathematical statistics, for small values of α, e.g., $\alpha = 5\%$ or $\alpha = 1\%$, is also used the notion *significance level of first type* or *type 1 error rate*. The most often used quantile for the most important distributions in practice are given in tables (**Table 21.16**, p. 1131, to **Table 21.20**, p. 1138).

16.2.2.3 Expected Value and Variance, Chebyshev Inequality

For a coarse characterization of the distribution of a random variable X, mostly the parameters *expected value*, denoted by μ, and *variance*, denoted by σ^2 are used. The expected value can be interpreted with the terminology of mechanics as the abscissa of the center of gravity of a surface bounded by the curve of the density function $f(x)$ and the x-axis. The variance represents a measure of deviation of the random variable X from its expected value μ.

1. Expected Value

If $g(X)$ is a function of the random variable X, then $g(X)$ is also a random variable. Its *expected value* or *expectation* is defined as:

a) Discrete Case: $$E(g(X)) = \sum_k g(x_k)p_k, \text{ if the series } \sum_{k=1}^{\infty} |g(x_k)|p_k \text{ exists.} \tag{16.49a}$$

b) Continuous Case: $$E(g(X)) = \int\limits_{-\infty}^{+\infty} g(x)f(x)\,dx, \text{ if } \int_{-\infty}^{+\infty} |g(x)|f(x)\,dx \text{ exists.} \tag{16.49b}$$

The expected value of the random variable with $g(X) = X$ is defined as

$$\mu_X = E(X) = \sum_k x_k p_k \quad \text{or} \quad \int\limits_{-\infty}^{+\infty} xf(x)\,dx, \tag{16.50a}$$

if the corresponding sum or integral with the absolute values exists. Because of (16.49a,b)

$$E(aX + b) = a\mu_X + b \quad (a, b \text{ const}) \tag{16.50b}$$

is also valid. Of course, it is possible that a random variable do not have any expected value.

2. Moments of Order n

Furthermore one introduces:

a) **Moment of Order n:** $E(X^n)$, (16.51a)

b) **Central Moment of Order n:** $E((X - \mu_X)^n)$. (16.51b)

3. Variance and Standard Deviation

In particular, for $n = 2$, the central moment is called the *variance* or *dispersion*:

$$E((X - \mu_X)^2) = D^2(X) = \sigma_X^2 = \begin{cases} \sum\limits_k (x_k - \mu_X)^2 p_k & \text{or} \\ \int\limits_{-\infty}^{+\infty} (x - \mu_X)^2 f(x)\,dx, \end{cases} \tag{16.52}$$

if the expected values occurring in the formula exist. The quantity σ_X is called the *standard deviation*. The following relations are valid:

$$D^2(X) = \sigma_X^2 = E(X^2) - \mu_X^2, \quad D^2(aX + b) = a^2 D^2(X). \tag{16.53}$$

4. Weighted and Arithmetical Mean

In the discrete case, the expected value is obviously the *weighted mean*

$$E(X) = p_1x_1 + \ldots + p_nx_n \tag{16.54}$$

of the values $x_1, \ldots, x_n$ with the probabilities p_k as *weights* ($k = 1, \ldots, n$). The probabilities for the uniform distribution are $p_1 = p_2 = \ldots = p_n = 1/n$, and $E(X)$ is the arithmetical mean of the values

x_k:

$$E(X) = \frac{x_1 + x_2 + \ldots + x_n}{n}. \tag{16.55}$$

In the continuous case, the density function of the continuous uniform distribution on the finite interval $[a, b]$ is

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \le x \le b, \\ 0 & \text{otherwise,} \end{cases} \tag{16.56}$$

and it follows that

$$E(X) = \frac{1}{b-a}\int_a^b x\,dx = \frac{a+b}{2}, \quad \sigma_X^2 = \frac{(b-a)^2}{12}. \tag{16.57}$$

5. Chebyshev Inequality

If the random variable X has the expected value μ and standard deviation σ, then for arbitrary $\lambda > 0$ the *Chebyshev inequality* (see 1.4.2.10, p. 31) is valid:

$$P(|X - \mu| \ge \lambda\sigma) \le \frac{1}{\lambda^2}. \tag{16.58}$$

That is, it is very unlikely that the values of the random variable X are farther from the expected value μ than a multiple of the standard deviation (λ large).

16.2.2.4 Multidimensional Random Variable

If the elementary events mean that n random variables $X_1, \ldots, X_n$ take n real values $x_1, \ldots, x_n$, then a *random vector* $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ is defined (see also random vector, 16.3.1.1, **4.**, p. 830). The corresponding distribution function is defined by

$$F(x_1, \ldots, x_n) = P(X_1 \le x_1, \ldots, X_n \le x_n). \tag{16.59}$$

The random vector is called continuous if there is a function $f(t_1, \ldots, t_n)$ such that

$$F(x_1, \ldots, x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f(t_1, \ldots, t_n)\,dt_1 \ldots dt_n \tag{16.60}$$

holds. The function $f(t_1, \ldots, t_n)$ is called the *density function*. It is non-negative. If some of the variables $x_1, \ldots, x_n$ tend to infinity, then one gets the so-called *marginal distributions*. Further investigations and examples can be found in the literature.

The random variables $X_1, \ldots, X_n$ are *independent random variables* if

$$F(x_1, \ldots, x_n) = F_1(x_1)F_2(x_2)\ldots F_n(x_n), \quad f(t_1, \ldots, t_n) = f_1(t_1)\ldots f_n(t_n). \tag{16.61}$$

16.2.3 Discrete Distributions

1. Two-Stage Population and Urn Model

Suppose one has a two-stage population with N elements, i.e., the population to be considered has two classes of elements. One class has M elements with a property A, the other one has $N - M$ elements which does not have the property A. If investigating the probabilities $P(A) = p$ and $P(\overline{A}) = 1 - p$ for randomly chosen elements, then one has to distinguish between two cases: When selecting n elements one after the other, either the previously selected element is replaced before selecting the next one, or it is not. The selected n elements, which contain k elements with the property A, is called the *sample*, n being the *size of the sample*. This can be illustrated by the urn model.

2. Urn Model

Suppose there are a lot of black balls and white balls in a container. The question is: What is the probability that among n randomly selected balls there are k black ones. If putting every chosen ball

back into the container after the determination of its color, then the number k of black ones among the chosen n balls has a *binomial distribution*. If one does not put back the chosen balls and $n \le M$ and $n \le N - M$, then the number of black ones has a *hypergeometric distribution*.

16.2.3.1 Binomial Distribution

Suppose in an experiment only the two events A and $\overline{A}$ are possible and the experiment is repeated n times and the accompanying probabilities are $P(A) = p$ and $P(\overline{A}) = 1 - p$ every time, then the probability that A takes place exactly k times is

$$W_p^n(k) = \binom{n}{k} p^k (1-p)^{n-k} \quad (k = 0, 1, 2, \ldots, n). \tag{16.62}$$

For every choice of an independent element from the population, the probabilities are

$$P(A) = \frac{M}{N} = p\,, \quad P(\overline{A}) = \frac{N-M}{N} = 1 - p = q. \tag{16.63}$$

The probability of getting an element with property A for the first k choices, then an element with the remaining property $\overline{A}$ for the $n - k$ choices is $p^k(1-p)^{n-k}$, because the results of choices are independent of each other. The sequence of the choices plays no role because the combinations have the same probability independently of the order of the choices, and these events are mutually exclusive, so one adds

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \tag{16.64}$$

equal numbers to get the required probability.
A random variable X_n, for which $P(X_n = k) = W_p^n(k)$ holds, is called *binomially distributed* with parameters n and p.

1. Expected Value and Variance

$$E(X_n) = \mu = n \cdot p, \tag{16.65a}$$

$$D^2(X_n) = \sigma^2 = n \cdot p(1-p). \tag{16.65b}$$

2. Approximation of the Binomial Distribution by the Normal Distribution

If X_n has a binomial distribution, then

$$\lim_{n\to\infty} P\left(\frac{X_n - E(X_n)}{D(X_n)} \le \lambda\right) = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{\lambda} \exp\left(\frac{-t^2}{2}\right)\, dt. \tag{16.65c}$$

This means that, if n is large, the binomial distribution can be well approximated by a normal distribution (see 16.2.4.1, p. 818) with parameters $\mu_X = E(X_n)$ and $\sigma^2 = D^2(X_n)$, if p or $1 - p$ are not too small. The approximation is the more accurate the closer p is to 0.5 and the larger n is, but acceptable if $np > 4$ and $n(1-p) > 4$ hold. For very small p or $1-p$, the approximation by the Poisson distribution (see (16.68) in 16.2.3.3) is useful.

3. Recursion Formula

The following recursion formula is recommended for practical calculations with the binomial distribution:

$$W_p^n(k+1) = \frac{n-k}{k+1} \cdot \frac{p}{q} \cdot W_p^n(k). \tag{16.65d}$$

4. Sum of Binomially Distributed Random Variables

If X_n and X_m are both binomially distributed random variables with parameters n, p and m, p, then the random variable $X = X_n + X_m$ is also binomially distributed with parameters $n + m$, p.

Fig. 16.2a,b,c represents the distributions of three binomially distributed random variables with parameters $n = 5$; $p = 0.5$, 0.25, and 0.1. Since the binomial coefficients are symmetric, the distribution

is symmetric for $p = q = 0.5$, and the farther p is from 0.5 the less symmetric the distribution is.

16.2.3.2 Hypergeometric Distribution

Just as with the binomial distribution, a two-stage population with N elements is considered, i.e., the population has two classes of elements. One class has M elements with a property A, the other one has $N - M$ elements which does not have the property A. In contrast to the case of binomial distribution, the chosen ball of the urn model is not replaced before choosing the next one.
The probability that among the n chosen balls there are k black ones is

$$P(X = k) = W^n_{M,N}(k) = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}} \quad \text{with} \tag{16.66a}$$

$$0 \le k \le n, \quad k \le M, \quad n - k \le N - M. \tag{16.66b}$$

If also $n \le M$ and $n \le N - M$ hold, then the random variable X with the distribution (16.66a) is called *hypergeometrically distributed*.

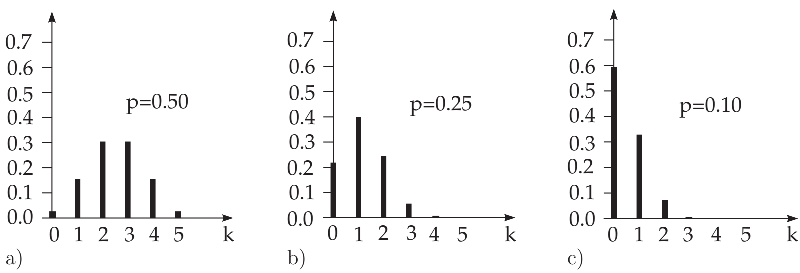


Figure 16.2

1. Expected Value and Variance of the Hypergeometric Distribution

$$\mu = E(X) = \sum_{k=0}^{n} k \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{k}} = n\frac{M}{N}, \tag{16.67a}$$

$$\sigma^2 = D^2(X) = E(X^2) - [E(X)]^2 = \sum_{k=0}^{n} k^2 \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{k}} - \left(n\frac{M}{N}\right)^2$$

$$= n\frac{M}{N}\left(1 - \frac{M}{N}\right)\frac{N-n}{N-1}. \tag{16.67b}$$

2. Recursion Formula

$$W^n_{M,N}(k+1) = \frac{(n-k)(M-k)}{(k+1)(N-M-n+k+1)} W^n_{M,N}(k). \tag{16.67c}$$

In **Fig. 16.3a,b,c** three hypergeometric distributions are represented for the cases $N = 100$, $M = 50$, 25 and 10, for $n = 5$. These cases correspond to the cases $p = 0.5$; 0.25, and 0.1 of **Fig. 16.2a,b,c**.

There is no significant difference between the binomial and hypergeometric distributions in these examples. If also M and $N - M$ are much larger than n, then the hypergeometric distribution can be well approximated by a binomial one with parameters as in (16.63).

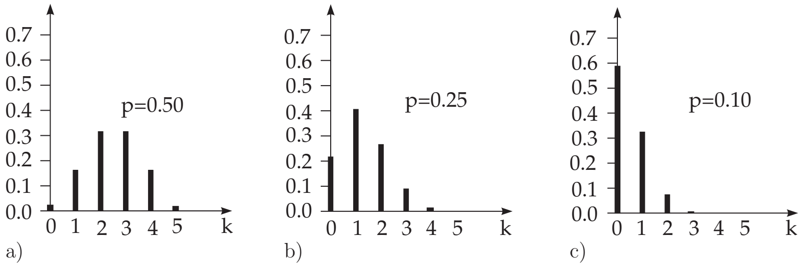


Figure 16.3

16.2.3.3 Poisson Distribution

If the possible values of a random variable X are the non-negative integers with probabilities

$$P(X = k) = \frac{\lambda^k}{k!}e^{-\lambda} \qquad (k = 0, 1, 2, \ldots; \ \lambda > 0), \tag{16.68}$$

then it has a *Poisson distribution* with parameter λ.

1. Expected Value and Variance of the Poisson Distribution

$$E(X) = \lambda, \quad (16.69a) \qquad D^2(X) = \lambda. \quad (16.69b)$$

2. Sum of Independent Poisson Distributed Random Variables

If X_1 and X_2 are independent Poisson distributed random variables with parameters λ_1 and λ_2, then the random variable $X = X_1 + X_2$ also has a Poisson distribution with parameter $\lambda = \lambda_1 + \lambda_2$.

3. Recursion Formula

$$P(X = k + 1) = \frac{\lambda}{k + 1}P(X = k)). \tag{16.69c}$$

4. Connection between Poisson and Binomial Distribution

The Poisson distribution can be obtained as a limit of binomial distributions with parameters n and p if $n \to \infty$, and p $(p \to 0)$ changes with n so that $np = \lambda =$ const, i.e., the Poisson distribution is a good approximation for a binomial distribution for large n and small p with $\lambda = np$. In practice, one uses it if $p \leq 0.08$ and $n \geq 1500p$ hold, because the calculations are easier with a Poisson distribution.

Table 21.16, p. 1131, contains numerical values for the Poisson distribution. **Fig. 16.4a,b,c** represents three Poisson distributions with $\lambda = np = 2.5$, 1.25 and 0.5, i.e., with parameters corresponding to **Figs. 16.2** and **16.3**.

5. Application

The number of independently occurring point-like discontinuities in a continuous medium can usually be described by a Poisson distribution, e.g., number of clients arriving in a store during a certain time interval; number of misprints in a book, the rate of radioactive decay, etc.

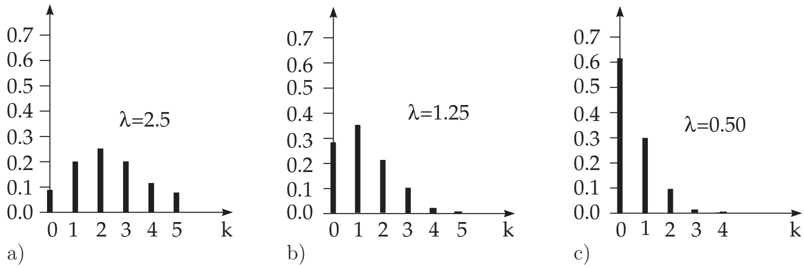


Figure 16.4

16.2.4 Continuous Distributions

16.2.4.1 Normal Distribution

1. Distribution Function and Density Function

A random variable X has a *normal distribution* if its distribution function is

$$P(X \le x) = F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int\limits_{-\infty}^{x} e^{-\frac{(t-\mu)^2}{2\sigma^2}}\, dt. \tag{16.70a}$$

Then it is also called a *normal variable*, and the distribution is called a (μ, σ^2) *normal distribution.* The function

$$f(t) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(t-\mu)^2}{2\sigma^2}} \tag{16.70b}$$

is the density function of the normal distribution. It takes its maximum at $t = \mu$ and it has inflection points at $\mu \pm \sigma$ (see (2.59), p. 73, and **Fig. 16.5a**).

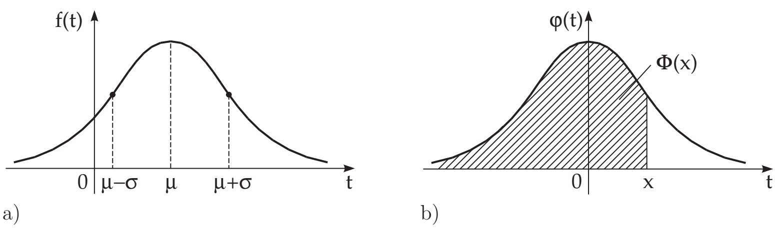


Figure 16.5

2. Expected Value and Variance

The parameters μ and σ^2 of the normal distribution are its expected value and variance, respectively, i.e.,

$$E(X) = \frac{1}{\sigma\sqrt{2\pi}} \int\limits_{-\infty}^{+\infty} x e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, dx = \mu, \tag{16.71a}$$

$$D^2(X) = E[(X-\mu)^2] = \frac{1}{\sigma\sqrt{2\pi}} \int\limits_{-\infty}^{+\infty} (x-\mu)^2 e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, dx = \sigma^2. \tag{16.71b}$$

If the normal random variables X_1 and X_2 are independent with parameters μ_1, σ_1 and μ_2, σ_2, resp., then the random variable $X = k_1X_1 + k_2X_2$ (k_1, k_2 real constants) also has a normal distribution with parameters $\mu = k_1\mu_1 + k_2\mu_2$, $\sigma = \sqrt{{k_1}^2{\sigma_1}^2 + {k_2}^2{\sigma_2}^2}$.

By the substitution $\tau = \dfrac{t-\mu}{\sigma}$ in (16.70a), the calculation of the values of the distribution function of any normal distribution is reduced to the calculation of the values of the distribution function of the $(0,1)$ normal distribution, which is called the *standard normal distribution*. Consequently, the probability $P(a \le X \le b)$ of a normal variable can be expressed by the distribution function $\Phi(x)$ of the standard normal distribution:

$$P(a \le X \le b) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right). \tag{16.72}$$

16.2.4.2 Standard Normal Distribution, Gaussian Error Function

1. Distribution Function and Density Function

From (16.70a) with $\mu = 0$ and $\sigma^2 = 1$ follows the distribution function

$$P(X \le x) = \Phi(x) = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{x} e^{-\frac{t^2}{2}}\, dt = \int\limits_{-\infty}^{x} \varphi(t)\, dt \tag{16.73a}$$

of the so-called *standard normal distribution*. Its density function is

$$\varphi(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}, \tag{16.73b}$$

it is called the *Gaussian error curve* **(Fig. 16.5b)**.

The values of the distribution function $\Phi(x)$ of the $(0,1)$ normal distribution are given in **Table 21.17**, p. 1133. Only the values for the positive arguments x are given, while the values for the negative arguments can be got from the relation

$$\Phi(-x) = 1 - \Phi(x). \tag{16.74}$$

2. Probability Integral

The integral $\Phi(x)$ is also called the *probability integral* or *Gaussian error integral*. In the literature the functions $\Phi_0(x)$ and erf (x) are sometimes denoted as error integral with the following definitions:

$$\Phi_0(x) = \frac{1}{\sqrt{2\pi}} \int\limits_{0}^{x} e^{-\frac{t^2}{2}}\, dt = \Phi(x) - \frac{1}{2}, \quad (16.75a) \qquad \operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int\limits_{0}^{x} e^{-t^2}\, dt = 2\cdot\Phi_0(\sqrt{2}x). \quad (16.75b)$$

With erf the error function is denoted.

16.2.4.3 Logarithmic Normal Distribution

1. Density Function and Distribution Function

The continuous random variable X has a *logarithmic normal distribution*, or *lognormal distribution* with parameters μ_L and σ_L^2 if it can take all positive values, and if the random variable Y, defined by

$$Y = \log X, \tag{16.76}$$

has a normal distribution with expected value μ_L and variance σ_L^2 (see also remark b) on page 820). Consequently, the random variable X has the density function

$$f(t) = \begin{cases} 0 & \text{for} \quad t \le 0, \\ \dfrac{\log e}{t\sigma_L\sqrt{2\pi}} \exp\left(-\dfrac{(\log t - \mu_L)^2}{2\sigma_L^2}\right) & \text{for} \quad t > 0, \end{cases} \tag{16.77a}$$

and the distribution function

$$F(x) = \begin{cases} 0 & \text{for} \quad x \le 0, \\ \dfrac{1}{\sigma_L\sqrt{2\pi}} \displaystyle\int_{-\infty}^{\log x} \exp\left(-\dfrac{(t-\mu_L)^2}{2\sigma_L^2}\right) dt & \text{for} \quad x > 0. \end{cases} \tag{16.77b}$$

In practical applications either the natural or the decimal logarithm is used.

2. Expected Value and Variance

Using the natural logarithm one gets the expected value and the variance of the lognormal distribution as:

$$\mu = \exp\left(\mu_L + \frac{\sigma_L^2}{2}\right), \qquad \sigma^2 = \left(\exp\sigma_L^2 - 1\right)\exp\left(2\mu_L + \sigma_L^2\right). \tag{16.78}$$

3. Remarks

a) The density function of the lognormal distribution is continuous everywhere and it has positive values only for positive arguments. **Fig. 16.6** shows the density functions of lognormal distributions for different μ_L and σ_L. Here has been used the natural logarithm.

b) Here the values μ_L and σ_L^2 are not the expected value and variance of the lognormal random variable itself, but of the variable $Y = \log X$, while μ and σ^2 are in conformity with (16.78) expected value and variance of the random variable X.

c) The values of the distribution function $F(x)$ of the lognormal distribution can be calculated by the distribution function $\Phi(x)$ of the standard normal distribution (see (16.73a)), in the following way:

$$F(x) = \Phi\left(\frac{\log x - \mu_L}{\sigma_L}\right). \tag{16.79}$$

d) The lognormal distribution is often applied in lifetime analysis of economical, technical, and biological processes.

e) The normal distribution can be used in additive superposition of a large number of independent random variables, and the lognormal distribution is used for multiplicative superposition of a large number of independent random variables.

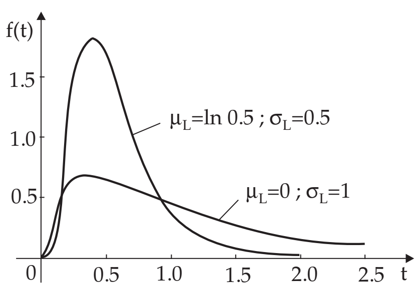


Figure 16.6

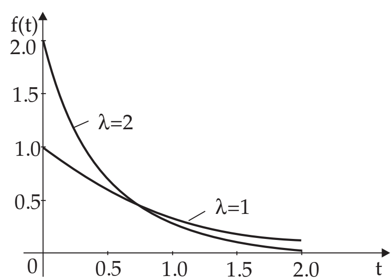


Figure 16.7

16.2.4.4 Exponential Distribution

1. Density Function and Distribution Function

A continuous random variable X has an *exponential distribution* with parameter λ ($\lambda > 0$) if its density function is (**Fig. 16.7**)

$$f(t) = \begin{cases} 0 & \text{for} \quad t < 0 \\ \lambda e^{-\lambda t} & \text{for} \quad t \ge 0, \end{cases} \tag{16.80a}$$

consequently, the distribution function is

$$F(x) = \int_{-\infty}^{x} f(t)dt = \begin{cases} 0 & \text{for} \quad x < 0, \\ 1 - e^{-\lambda x} & \text{for} \quad x \ge 0. \end{cases} \tag{16.80b}$$

2. Expected Value and Variance

$$\mu = \frac{1}{\lambda}, \qquad \sigma^2 = \frac{1}{\lambda^2}. \tag{16.81}$$

Usually, the following quantities are described by an exponential distribution: Length of phone calls, lifetime of radioactive particles, working time of a machine between two stops in certain processes, lifetime of light-bulbs or certain building elements.

16.2.4.5 Weibull Distribution

1. Density Function and Distribution Function

The continuous random variable X has a Weibull distribution with parameters α and β ($\alpha > 0, \beta > 0$), if its density function is

$$f(t) = \begin{cases} 0 & \text{for} \quad t < 0, \\ \frac{\alpha}{\beta}\left(\frac{t}{\beta}\right)^{\alpha-1} \exp\left[-\left(\frac{t}{\beta}\right)^{\alpha}\right] & \text{for} \quad t \ge 0 \end{cases} \tag{16.82a}$$

and so its distribution function is

$$F(x) = \begin{cases} 0 & \text{for} \quad x < 0, \\ 1 - \exp\left[-\left(\frac{x}{\beta}\right)^{\alpha}\right] & \text{for} \quad x \ge 0. \end{cases} \tag{16.82b}$$

2. Expected Value and Variance

$$\mu = \beta\Gamma\left(1 + \frac{1}{\alpha}\right), \qquad \sigma^2 = \beta^2\left[\Gamma\left(1 + \frac{2}{\alpha}\right) - \Gamma^2\left(1 + \frac{1}{\alpha}\right)\right]. \tag{16.83}$$

Here $\Gamma(x)$ denotes the gamma function (see 8.2.5, **6.**, p. 514):

$$\Gamma(x) = \int_0^{\infty} t^{x-1}e^{-t}\,dt \quad \text{for} \quad x > 0. \tag{16.84}$$

In (16.82a), α is the shape parameter and β is the scale parameter **(Fig. 16.8, Fig. 16.9)**.

Remarks:

a) The Weibull distribution becomes an exponential distribution for $\alpha = 1$ with $\lambda = \frac{1}{\beta}$.

b) The Weibull distribution also has a three-parameter form by introducing a position parameter γ. Then the distribution function is:

$$F(x) = 1 - \exp\left[-\left(\frac{x-\gamma}{\beta}\right)^{\alpha}\right]. \tag{16.85}$$

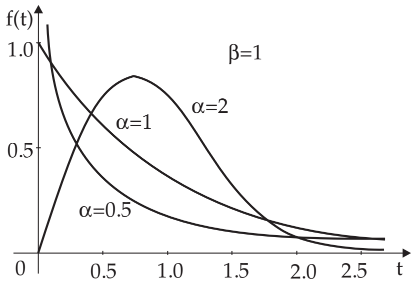


Figure 16.8

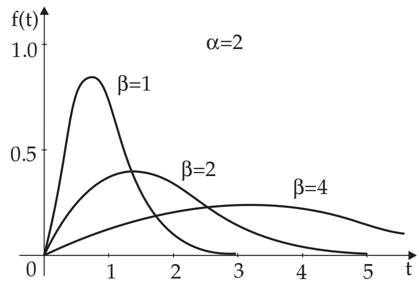


Figure 16.9

c) The Weibull distribution is especially useful in life expectancy theory, because, e.g., it describes the functional lifetime of building elements with great flexibility.

16.2.4.6 χ^2 (Chi-Square) Distribution

1. Density Function and Distribution Function

Let X_1, X_2,..., X_n be n independent $(0,1)$ normal random variables. Then the distribution of the random variable

$$\chi^2 = {X_1}^2 + {X_2}^2 + \cdots + {X_n}^2 \tag{16.86}$$

is called the χ^2 *distribution* with n degrees of freedom. Its distribution function is denoted by $F_{\chi^2}(x)$, and the corresponding density function by $f_{\chi^2}(t)$.

$$f_{\chi^2}(t) = \begin{cases} \dfrac{1}{2^{n/2}\Gamma\left(\dfrac{n}{2}\right)} t^{\frac{n}{2}-1} e^{-\frac{t}{2}} & \text{for } (t > 0) \\ 0 & \text{for } t \le 0. \end{cases} \tag{16.87a}$$

$$F_{\chi^2}(x) = P(\chi^2 \le x) = \frac{1}{2^{n/2}\Gamma\left(\dfrac{n}{2}\right)} \int_0^x t^{\frac{n}{2}-1} e^{-\frac{t}{2}}\, dt \quad (x > 0). \tag{16.87b}$$

2. Expected Value and Variance

$$E(\chi^2) = n, \tag{16.88a}$$

$$D^2(\chi^2) = 2n. \tag{16.88b}$$

3. Sum of Independent Random Variables

If X_1 and X_2 are independent random variables both having a χ^2 distribution with n and m degrees of freedom, then the random variable $X = X_1 + X_2$ has a χ^2 distribution with $n+m$ degrees of freedom.

4. Sum of Independent Normal Random Variables

If X_1, X_2,..., X_n are independent, $(0,\sigma)$ normal random variables, then

$$X = \sum_{i=1}^{n} {X_i}^2 \quad \text{has the density function} \quad f(t) = \frac{1}{\sigma^2} f_{\chi^2}\left(\frac{t}{\sigma^2}\right), \tag{16.89}$$

$$X = \frac{1}{n}\sum_{i=1}^{n} {X_i}^2 \quad \text{has the density function} \quad f(t) = \frac{n}{\sigma^2} f_{\chi^2}\left(\frac{nt}{\sigma^2}\right), \tag{16.90}$$

$$X = \sqrt{\frac{1}{n}\sum_{i=1}^{n} {X_i}^2} \quad \text{has the density function} \quad f(t) = \frac{2t}{\sigma^2} f_{\chi^2}\left(\frac{t^2}{\sigma^2}\right). \tag{16.91}$$

5. Quantile

For the quantile (see 16.2.2.2, **3.**, p. 812) $\chi^2_{\alpha,m}$ of the χ^2 distribution with m degrees of freedom (**Fig. 16.10**),

$$P(X > \chi^2_{\alpha,m}) = \alpha. \tag{16.92}$$

Quantiles of the χ^2 distribution can be found in **Table 21.18**, p. 1135.

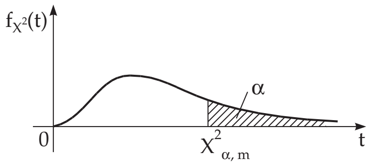


Figure 16.10

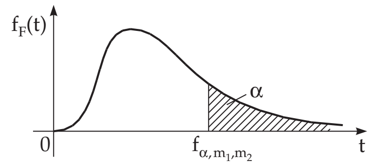


Figure 16.11

16.2.4.7 Fisher *F* Distribution

1. Density Function and Distribution Function

If X_1 and X_2 are independent random variables both having χ^2 distribution with m_1 and m_2 degrees of freedom, then the distribution of the random variable

$$F_{m_1,m_2} = \frac{X_1}{m_1} \Big/ \frac{X_2}{m_2} \tag{16.93}$$

is a Fisher distribution or F *distribution* with m_1, m_2 degrees of freedom. The density function is

$$f_F(t) = \begin{cases} \left(\frac{m_1}{2}\right)^{m_1/2} \left(\frac{m_2}{2}\right)^{m_2/2} \frac{\Gamma\left(\frac{m_1}{2} + \frac{m_2}{2}\right)}{\Gamma\left(\frac{m_1}{2}\right)\Gamma\left(\frac{m_2}{2}\right)} \frac{t^{\frac{m_1}{2}} - 1}{\left(\frac{m_1}{2}t + \frac{m_2}{2}\right)^{\frac{m_1}{2} + \frac{m_2}{2}}} & \text{for } t > 0, \\ 0 & \text{for } t \le 0. \end{cases} \tag{16.94a}$$

For $x \le 0$ holds $F_F(x) = P(F_{m_1,m_2} \le x) = 0$, for $x > 0$:

$$\begin{aligned} F_F(x) &= P(F_{m_1,m_2} \le x) \\ &= \left(\frac{m_1}{2}\right)^{m_1/2} \left(\frac{m_2}{2}\right)^{m_2/2} \frac{\Gamma\left(\frac{m_1}{2} + \frac{m_2}{2}\right)}{\Gamma\left(\frac{m_1}{2}\right)\Gamma\left(\frac{m_2}{2}\right)} \int_0^x \frac{\left(t^{\frac{m_1}{2}} - 1\right) dt}{\left(\frac{m_1}{2}t + \frac{m_2}{2}\right)^{\frac{m_1}{2} + \frac{m_2}{2}}} \end{aligned} \tag{16.94b}$$

2. Expected Value and Variance

$$E(F_{m_1,m_2}) = \frac{m_2}{m_2 - 2}, \tag{16.95a}$$

$$D^2(F_{m_1,m_2}) = \frac{2{m_2}^2(m_1 + m_2 - 2)}{m_1(m_2 - 2)^2(m_2 - 4)}. \tag{16.95b}$$

3. Quantile

The quantiles (see 16.2.2.2, **3.**, p. 812) t_{α,m_1,m_2} of the Fisher distribution (**Fig. 16.11**) can be found in **Table 21.19**, p. 1136.

16.2.4.8 Student t Distribution

1. Density Function and Distribution Function

If X is a $(0,1)$ normal random variable and Y is a random variable independent from X and it has a χ^2 distribution with $m = n-1$ degrees of freedom, then the distribution of the random variable

$$T = \frac{X}{\sqrt{Y/m}} \tag{16.96}$$

is called a *Student t distribution* or *t distribution* with m degrees of freedom. The distribution function is denoted by $F_S(x)$, and the corresponding density function by $f_S(t)$.

$$f_S(t) = \frac{1}{\sqrt{m\pi}} \frac{\Gamma\left(\frac{m+1}{2}\right)}{\Gamma\left(\frac{m}{2}\right)} \frac{1}{\left(1+\frac{t^2}{m}\right)^{\frac{m+1}{2}}}, \tag{16.97a}$$

$$F_S(x) = P(T \le x) = \int_{-\infty}^{x} f_S(t)\,dt = \frac{1}{\sqrt{m\pi}} \frac{\Gamma\left(\frac{m+1}{2}\right)}{\Gamma\left(\frac{m}{2}\right)} \int_{-\infty}^{x} \frac{dt}{\left(1+\frac{t^2}{m}\right)^{\frac{m+1}{2}}}. \tag{16.97b}$$

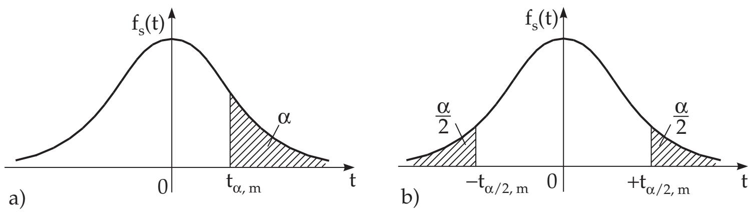


Figure 16.12

2. Expected Value and Variance

$$E(T) = 0 \quad (m > 1), \tag{16.98a}$$

$$D^2(T) = \frac{m}{m-2} \quad (m > 2). \tag{16.98b}$$

3. Quantile

The quantiles $t_{\alpha,m}$ and $t_{\alpha/2,m}$ of the t distribution (**Fig. 16.12a,b**), for which

$$P(T > t_{\alpha,m}) = \alpha \tag{16.99a}$$

or

$$P(|T| > t_{\alpha/2,m}) = \alpha \tag{16.99b}$$

holds, are given in **Table 21.20**, p. 1138.

The Student t distribution, introduced by Gosset under the name Student, is used in the case of samples with small sample size n, when only estimations can be given for the mean and for the standard deviation. The standard deviation (16.98b) no longer depends on the deviation of the population from where the sample is taken.

16.2.5 Law of Large Numbers, Limit Theorems

The law of large numbers gives a relation between the probability $P(A)$ of a random event A and its relative frequency n_A/n with a large number of repeated experiments.

1. Law of Large Numbers of Bernoulli

The following inequality holds for arbitrary given numbers $\varepsilon > 0$ and $\eta > 0$

$$P\left(\left|\frac{n_A}{n} - P(A)\right| < \varepsilon\right) \geq 1 - \eta, \quad (16.100a) \quad \text{if} \quad n \geq \frac{1}{4\varepsilon^2\eta}. \quad (16.100b)$$

For other similar theorems see [16.5].

■ How many times should a not necessarily fair die be rolled if the relative frequency of the 6 should be closer to its probability than 0.1 with a probability of at least 95 % ?
Now, $\varepsilon = 0.01$ and $\eta = 0.05$, so $4\varepsilon^2\eta = 2 \cdot 10^{-5}$, and according to the law of large numbers of Bernoulli $n \geq 5 \cdot 10^4$ must hold. This is an extremely large number, which can be reduced, if the distribution function is known (see [16.12]).

2. Central Limit Theorem of Lindeberg–Levy

If the independent random variables $X_1, \ldots, X_n$ all have the same distribution with an expected value μ and a variance σ^2, then the distribution of the random variable

$$Y_n = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i - \mu}{\sigma/\sqrt{n}} \quad (16.101)$$

tends to the $(0,1)$ normal distribution for $n \to \infty$, i.e., for its distribution function $F_n(y)$ follows

$$\lim_{n\to\infty} F_n(y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y} e^{-\frac{t^2}{2}}\, dt. \quad (16.102)$$

If $n > 30$ holds, then $F_n(y)$ can be replaced by the $(0,1)$ normal distribution. Further limit theorems can be found in [16.5], [16.7].

■ Given a sample of 100 items from a production of resistors. It is supposed that their actual resistance values are independent and they have the same distribution with deviation $\sigma^2 = 150$. The mean value for these 100 resistors is $\overline{x} = 1050\,\Omega$. In which domain is the true expected value μ with a probability of 99 % ?
Looking for an ε such that $P(|\overline{X} - \mu| \leq \varepsilon) = 0.99$ holds. Supposing (see (16.101)) that the random variable $Y = \dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}}$ has a (0.1) normal distribution. From $P(|Y| \leq \lambda) = P(-\lambda \leq Y \leq \lambda) = P(Y \leq \lambda) - P(Y < -\lambda)$, and from $P(Y \leq -\lambda) = 1 - P(Y \leq \lambda)$ it follows that $P(|Y| \leq \lambda) = 2P(Y \leq \lambda) - 1 = 0.99$.
So, $P(Y \leq \lambda) = \Phi(\lambda) = 0.995$ and from **Table 21.17**, p. 1133, one gets $\lambda = 2.58$. Since $\sigma/\sqrt{100} = 1.225$ there is with a 99 % probability: $|1050 - \mu| < 2.58 \cdot 1.225$, i.e., $1046.8\,\Omega < \mu < 1053.2\,\Omega$.

16.2.6 Stochastic Processes and Stochastic Chains

Many processes occurring in nature and those being studied in engineering and economics can be realistically described only by time-dependent random variables.

■ The electric consumption of a city at a certain time t has a random fluctuation that is dependent on the actual demand of the households and industry. The electric consumption can be considered as a continuous random variable X. When the observation time t changes, electric consumption is a continuous random variable at every moment, so it is a function of time.

The stochastic analysis of time-dependent random variables leads to the concept of stochastic pro-

cesses, which has a huge literature of its own (see, e.g., [16.7], [16.9]). Some introductory notions will be given next.

16.2.6.1 Basic Notions, Markov Chains

1. Stochastic Processes

A set of random variables depending on one parameter is called a *stochastic process*. The parameter, in general, can be considered as time t, so the random variable can be denoted by X_t and the stochastic process is given by the set

$$\{X_t | t \in T\}. \tag{16.103}$$

The set of parameter values is called the *parameter space* T, the set of values of the random variables is the *state space* Z.

2. Stochastic Chains

If both the parameter space and the state space are discrete, i.e., the state variable X_t and the parameter t can have only finite or countably infinite different values, then the stochastic process is called a *stochastic chain*. In this case the different states and different parameter values can be numbered:

$$Z = \{1, 2, \ldots, i, i+1, \ldots, \} \tag{16.104}$$

$$T = \{t_0, t_1, \ldots, t_m, t_{m+1}, \ldots\} \text{ with } 0 \le t_0 < t_1 < \ldots < t_m < t_{m+1} < \ldots . \tag{16.105}$$

The times $t_0, t_1, \ldots$ are not necessary equally spaced.

3. Markov Chains, Transition Probabilities

If the probability of the different values of $X_{t_{m+1}}$ in a stochastic process depends only on the state at time t_m, then the process is called a *Markov chain*. The Markov property is defined precisely by the requirement that

$$P(X_{t_{m+1}} = i_{m+1} | X_{t_0} = i_0, X_{t_1} = i_1, \ldots, X_{t_m} = i_m) = P(X_{t_{m+1}} = i_{m+1} | X_{t_m} = i_m)$$
$$\text{for all } m \in \{0, 1, 2, \ldots\} \text{ and for all } i_0, i_1, \ldots, i_{m+1} \in Z. \tag{16.106}$$

Consider a Markov chain and times t_m and t_{m+1}. The conditional probabilities

$$P(X_{t_{m+1}} = j | X_{t_m} = i) = p_{ij}(t_m, t_{m+1}) \tag{16.107}$$

are called the *transition probabilities* of the chain. The transition probability determines the probability by which the system changes from the state $X_{t_m} = i$ at t_m into the state $X_{t_{m+1}} = j$ at t_{m+1}.

If the state space of a Markov chain is finite, i.e., $Z = \{1, 2, \ldots, N\}$, then the transition probabilities $p_{ij}(t_1, t_2)$ between the states at times t_1 and t_2 can be represented by a quadratic matrix $\mathbf{P}(t_1, t_2)$, by the so-called *transition matrix*:

$$\mathbf{P}(t_1, t_2) = \begin{pmatrix} p_{11}(t_1, t_2) & p_{12}(t_1, t_2) & \ldots & p_{1N}(t_1, t_2) \\ p_{21}(t_1, t_2) & p_{22}(t_1, t_2) & \ldots & p_{2N}(t_1, t_2) \\ \vdots & & & \\ p_{N1}(t_1, t_2) & p_{N2}(t_1, t_2) & \ldots & p_{NN}(t_1, t_2) \end{pmatrix}. \tag{16.108}$$

The times t_1 and t_2 are not necessarily consecutive.

4. Time-Homogeneous (Stationary) Markov Chains

If the transition probabilities of a Markov chain (16.107) do not depend on time, i.e.,

$$p_{ij}(t_m, t_{m+1}) = p_{ij}, \tag{16.109}$$

then the Markov chain is called *time-homogeneous* or *stationary*. A stationary Markov chain with a finite state space $Z = \{1, 2, \ldots, N\}$ has the transition matrix

$$\mathbf{P} = \begin{pmatrix} p_{11} & p_{12} & \ldots & p_{1N} \\ p_{21} & p_{22} & \ldots & p_{2N} \\ \vdots & & & \\ p_{N1} & p_{N2} & \ldots & p_{NN} \end{pmatrix}, \tag{16.110a}$$

where

$$\text{a)}\quad p_{ij} \geq 0 \text{ for all } i, j \text{ and} \tag{16.110b}$$

$$\text{b)}\quad \sum_{j=1}^{N} p_{ij} = 1 \text{ for all } i. \tag{16.110c}$$

Being independent of time p_{ij} gives the transition probability from the state i into the state j during time unit.

■ The number of busy lines in a telephone exchange can be modeled by a stationary Markov chain. For the sake of simplicity it is supposed that there are only two lines. Hence, the states are $i = 0, 1, 2$. Let the time unit be, e.g., 1 minute. Suppose the transition matrix p_{ij} is:

$$(p_{ij}) = \begin{pmatrix} 0.7 & 0.3 & 0.0 \\ 0.2 & 0.5 & 0.3 \\ 0.1 & 0.4 & 0.5 \end{pmatrix} \quad (i, j = 0, 1, 2).$$

In the matrix (p_{ij}) the first row corresponds to the state $i = 0$. The matrix element $p_{12} = 0,3$ (second row, third column) shows the probability that two lines are busy at time t_m given that one was busy at t_{m-1}.

Remark: Every quadratic matrix $\mathbf{P} = (p_{ij})$ of size $N \times N$ satisfying the properties (16.110b) and (16.110c) is called a *stochastic matrix*. Their row vectors are called *stochastic vectors*.

Although the transition probabilities of a stationary Markov chain do not depend on time, the distribution of the random variable X_t is given at a given time by the probabilities

$$P(X_t = i) = p_i(t) \quad (i = 1, 2, \ldots, N) \quad (16.111\text{a}) \qquad \text{with } \sum_{i=1}^{N} p_i(t) = 1 \tag{16.111b}$$

since the process is in one of the states with probability one at any time t.

5. Probability Vector and Transition Matrix

Probabilities (16.111a) can be written in the form of a *probability vector*

$$\underline{\mathbf{p}} = (p_1(t), p_2(t), \ldots, p_N(t)). \tag{16.112}$$

The probability vector $\underline{\mathbf{p}}$ is a stochastic vector. It determines the distribution of the states of a stationary Markov chain at time period t. Let the transition matrix $\mathbf{P}$ of a stationary Markov chain be given (according to (16.110a,b,c)). Starting with the probability distribution at time period t determine the probability distribution at $t + 1$, that is, calculate $\underline{\mathbf{p}}(t + 1)$ from $\mathbf{P}$ and $\underline{\mathbf{p}}(t)$:

$$\underline{\mathbf{p}}(t + 1) = \underline{\mathbf{p}}(t) \cdot \mathbf{P} \tag{16.113}$$

and furthermore

$$\underline{\mathbf{p}}(t + k) = \underline{\mathbf{p}}(t) \cdot \mathbf{P}^k. \tag{16.114}$$

Remarks:

1. For $t = 0$ it follows from (16.114) that

$$\underline{\mathbf{p}}(k) = \underline{\mathbf{p}}(0)\mathbf{P}^k, \tag{16.115}$$

that is, a stationary Markov chain is uniquely determined by the initial distribution $\underline{\mathbf{p}}(0)$ and the transition matrix $\mathbf{P}$.

2. If matrices $\mathbf{A}$ and $\mathbf{B}$ are stochastic matrices, then $\mathbf{C} = \mathbf{AB}$ is a stochastic matrix, as well. Consequently, if $\mathbf{P}$ is a stochastic matrix, then the powers $\mathbf{P}^k$ are also stochastic matrices.

■ A particle changes its position (state) X_t $(1 \leq x \leq 5)$ along a line in time periods $t = 1, 2, 3, \ldots$ according to the following rules:

a) If the particle is at $x = 2, 3, 4$, then it moves to the right by a unit during the next time unit with probability $p = 0.6$ and to the left with probability $1 - p = 0.4$.

b) At points $x = 1$ and $x = 5$ the particle is absorbed, i.e., it stays there with probability 1.
c) At time $t = 0$ the position of the particle is $x = 2$.
Determine the probability distribution $\underline{\mathbf{p}}(3)$ at time period $t = 3$.

By (16.115) the probability distribution $\underline{\mathbf{p}}(3) = \underline{\mathbf{p}}(0) \cdot \mathbf{P}^3$ holds with $\underline{\mathbf{p}}(0) = (0, 1, 0, 0, 0)$ and with the transition matrix

$$\mathbf{P} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0.4 & 0 & 0.6 & 0 & 0 \\ 0 & 0.4 & 0 & 0.6 & 0 \\ 0 & 0 & 0.4 & 0 & 0.6 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}. \text{ Hence } \mathbf{P}^3 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0.496 & 0 & 0.288 & 0 & 0.216 \\ 0.160 & 0.192 & 0 & 0.288 & 0.360 \\ 0.064 & 0 & 0.192 & 0 & 0.744 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

and finally $\underline{\mathbf{p}}(3) = (0.496; 0; 0.288; 0; 0.216)$.

16.2.6.2 Poisson Process

1. The Poisson Process

In the case of a stochastic chain both the state space Z and the parameter space T are discrete, that is, the stochastic process is observed only at discrete time periods $t_0, t_1, t_2, \ldots$. Now, there is studied a process with continuous parameter space T, and it is called a *Poisson process.*

1. Mathematical Formulation of the Poisson Process The following assumptions are made for the mathematical formulation of the Poisson process:

a) Let the random variable X_t be the number of signals in the time interval [0,t);

b) Let the probability $p_X(t) = P(X_t = x)$ be the probability of x signals during the time interval $[0, t)$.

Additionally, the following assumptions are required, which hold in the process of radioactive decay and many other random processes (at least approximately):

c) The probability $P(X_t = x)$ of x signals in a time interval of length t depends only on x and t, and does not depend on the position of the time interval on the time axis.

d) The numbers of signals in disjoint time intervals are independent random variables.

e) The probability to get at least one signal in a very short interval of length Δt is approximately proportional to this length. The proportionality factor is denoted by λ $(\lambda > 0)$.

2. Distribution Function By properties **a)**–**e)** the distribution of the random variable X_t is determined. One gets:

$$P(X_t = x) = \frac{(\lambda t)^x}{x\,!} e^{-\lambda t}, \tag{16.116}$$

where $\mu = \lambda t$ is the expected value and $\sigma^2 = \lambda t$ the variance.

3. Remarks

1. From (16.116) the Poisson distribution follows as a special case for $t = 1$ (see 16.2.3.3, p.817).

2. To interpret the parameter λ or to estimate its value from observed data the following properties are useful:

- λ is the average number of signals during a time unit,
- $\dfrac{1}{\lambda}$ is the average distance (in time) between two signals in a Poisson process.

3. The Poisson process can be interpreted as the random motion of a particle in the state space $Z = \{0, 1, 2, \ldots\}$. The particle starts in the state 0, and at every sign it jumps from state i into the next state $i + 1$. Furthermore, for a small interval Δt the transition probability $p_{i,i+1}$ from state i into the state $i + 1$ should be:

$$p_{i,i+1} \approx \lambda \Delta t. \tag{16.117}$$

λ is called the transition rate.

4. Examples of Poisson Processes

■ Radioactive decay is a typical example of a Poisson process: The number of decays (signals) are registered with a counter and marked on the time axis. The observation interval should be relatively small with respect to the half-period of the radiating matter.

■ Consider the number of calls registered in a telephone exchange until time t and calculate, e.g., the probability that at most x calls are registered until time t with the assumption that the average number of calls during a time unit is λ.

■ In reliability testing, the number of failures of a reparable system is counted during a period of duty.

■ Queuing theory considers the number of customers arriving at the counter of a department store, to a booking office or to a gasoline station.

2. Birth and Death Processes

One of the generalizations of the Poisson process is the assumption that the transition rate λ_i in (16.117) depends on the state i. Another generalization is the transition from state i into state $i-1$ is allowed. The corresponding transition rate is denoted by μ_i. The state i can be considered, e.g., as the number of individuals in a population. It increases by one at transition from state i into state $i+1$, and decreases by one at transition from i into $i-1$. These stochastic processes are called *birth* and *death processes*. Let $p(X_t = i) = p_i(t)$ be the probability that the process is in state i at time t. Analogously to the Poisson process for the transition probability holds:

$$\begin{aligned} &\text{from } \; i-1 \; \text{ into } \; i: \; p_{i-1,i} \approx \lambda_{i-1}\Delta t,\\ &\text{from } \; i+1 \; \text{ into } \; i: \; p_{i+1,i} \approx \mu_{i+1}\Delta t,\\ &\text{from } \; i \; \text{ into } \; i: \; p_{i,i} \approx 1-(\lambda_i+\mu_i)\Delta t. \end{aligned} \tag{16.118}$$

Remark: The Poisson process is a pure birth process with a constant transition rate.

3. Queuing

The simplest queuing system is considered as a counter where customers are served one by one in the order of their arrival time. The waiting room is sufficiently large, so no one needs to leave because it becomes full. The customers arrive according to a Poisson process, that is, the inter-arrival time between two clients is exponentially distributed with parameter λ, and these inter-arrival times are independent. In many cases also the serving time has an exponential distribution with parameter μ. The parameters λ and μ have the following meanings:

- λ: average number of arrivals per time unit,
- $\dfrac{1}{\lambda}$: average inter-arrival time,
- μ: average number of served clients per time unit,
- $\dfrac{1}{\mu}$: average serving time.

Remarks:

1. If the number of clients standing in the queue is considered as the state of this stochastic process, then the above simple queuing model is a birth and death process with constant birth rate λ and constant death rate μ.

2. The above queuing model can be modified and generalized in many different ways, e.g., there can be several counters where the clients are served and/or the arrival times and serving times follow different distributions (see [16.9], [16.19]).

16.3 Mathematical Statistics

Mathematical statistics provides an application of probability theory for given mass phenomena. Its theorems allow to make statements with certain probability about properties of given sets, which statements are based on the results of experiments whose number should be kept low for economical reasons.

16.3.1 Statistic Function or Sample Function

16.3.1.1 Population, Sample, Random Vector

1. Population

The *Population* is the set of all elements of interest in a particular study. Any set of things having the same property in a certain sense can be considered, e.g., every article of a certain production process or all the values of a measuring sequence occurring in a permanent repetition of an experiment. The number N of the elements of a population can be very large, even practically infinite. The word *population* is often used to denote also the set of numerical values assigned to the elements.

2. Sample

In order not to check the total population about the considered property, data are collected only from a subset, from a so-called *sample* of size n $(n \le N)$. Talking about a random choice means that every element of the population has the same chance of being chosen. A *random sample* of size n from a finite population of size N is a sample selected such that each possible sample of size n has the same probability of being selected. A *random sample* from an infinite population is a sample selected such that each element is selected independently. The random choice can be made by mixing, blind taking out or by so-called *random numbers*. The word *sample* is used for the set of values assigned to the selected elements.

3. Random Choice with Random Numbers

It often happens that a random selection is physically impossible on the spot, e.g., in the case of piled material, like concrete stabs. Then random numbers are applied for a random selection (see **Table 21.21**, p. 1139).

Most calculators can generate uniformly distributed random numbers from the interval $[0, 1]$. Pressing the key RAN yields a number between $0.00\ldots0$ and $0.99\ldots9$. The digits after the decimal point form a sequence of random numbers.

Random numbers are often taken from tables. Two-digit random numbers are given in **Table 21.21**, p. 1139. If larger ones are necessary, then one can compose several-digit numbers by writing them after each other.

■ A random sample is to be examined from a transport of 70 piled pipes. The sample size is supposed to be 10. First the pipes are numbered from 00 to 69. A two-digit table of random numbers is applied to select the numbers. Then the way is fixed how to choose the numbers, e.g., horizontally, vertically or diagonally. If during this process random numbers occur repeatedly, or they are larger than 69, then they are simply omitted. The pipes corresponding to the chosen random numbers are the elements of the sample. If there is a several-digit table of random numbers, they can be decomposed into two-digit numbers.

4. Random Vector

A random variable X can be characterized by its distribution function, by its parameters, where the distribution function itself is determined completely by the properties of the population. These are unknown at the beginning of a statistical investigation, so one want to collect as much information as possible with the help of samples. Usually the investigation is not restricted to one sample but more samples are applied (with same size n if it is possible, for practical reasons). The elements of a sample are chosen randomly, so the realizations take their values randomly, i.e., the first value of the first sample is usually different from the first value of the second sample. Consequently, the first value of a sample is a random variable itself denoted by X_1. Analogously, the random variables $X_2, X_3, \ldots X_n$

can be introduced for the second, third,. . . , n-th sample values, and they are called *sample variables*. Together, they form the *random vector*

$$\underline{\mathbf{X}} = (X_1, X_2, \ldots, X_n). \tag{16.119a}$$

Every sample of size n with elements x_i can be considered as a vector

$$\underline{\mathbf{x}} = (x_1, x_2, \ldots, x_n), \tag{16.119b}$$

as a realization of the random vector.

16.3.1.2 Statistic Function or Sample Function

Since the samples are different from each other, their arithmetic means $\overline{x}$ are also different. They can be considered as realizations of a new random variable denoted by $\overline{X}$ which depends on the *sample variables* $X_1, X_2, \ldots, X_n$.

$$\begin{array}{rlllll} \text{1. sample:} & x_{11}, & x_{12}, & \ldots & x_{1n} & \text{with mean } \overline{x}_1. \\ \text{2. sample:} & x_{21}, & x_{22}, & \ldots & x_{2n} & \text{with mean } \overline{x}_2. \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ m\text{-th sample:} & x_{m1}, & x_{m2}, & \ldots & x_{mn} & \text{with mean } \overline{x}_m. \end{array} \tag{16.120}$$

The realization of the j-th sample variable in the i-th sample is denoted by x_{ij} $(i = 1, 2, \ldots, m;\ j = 1, 2, \ldots, n)$.

A function of the random vector $\underline{\mathbf{X}} = (X_1, X_2, \ldots, X_n)$ is again a random variable, and it is called a *statistic* or *sample function*. The most important sample functions are the mean, variance, median and range.

1. Mean

The mean $\overline{X}$ of the random variables X_i is:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{16.121a}$$

The mean $\overline{x}$ of the sample $(x_1, x_2, \ldots,\ x_n)$ is

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{16.121b}$$

It is often useful to introduce an *estimate value* x_0 in the calculations of the mean. It can be chosen arbitrarily but possibly close to the mean $\overline{x}$. If, e.g., x_i, $(i = 1, 2, \ldots)$ are several-digit numbers in a long measuring sequence, and they differ only in the last few digits, it is simpler to do the calculations only with the smaller numbers

$$z_i = x_i - x_0. \tag{16.121c}$$

Then follows

$$\overline{x} = x_0 + \frac{1}{n}\sum_{i=1}^{n} z_i = x_0 + \overline{z}. \tag{16.121d}$$

2. Variance

The variance S^2 of the random variables X_i with mean $\overline{X}$ is defined by:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2. \tag{16.122a}$$

The realization of the variance with the help of the sample $(x_1, x_2, \ldots, x_n)$ is

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2. \tag{16.122b}$$

It is proven that the estimation of the variance of the original population gives a more accurate estimation by dividing $n-1$ than by dividing n. With the estimated value x_0 follows

$$s^2 = \frac{\sum_{i=1}^{n} {z_i}^2 - \overline{z}\sum_{i=1}^{n} z_i}{n-1} = \frac{\sum_{i=1}^{n} {z_i}^2 - n(\overline{x}-x_0)^2}{n-1}. \tag{16.122c}$$

For $x_0 = \overline{x}$ the correction is $\overline{z}\sum_{i=1}^{n} z_i = 0$ because $\overline{z} = 0$ holds.

3. Median

Let the n elements of the sample be arranged in ascending (or descending) order. If n is odd, then the *median* $\tilde{X}$ is the value of the $\frac{n+1}{2}$-th item; if n is even, then the *median* is the average value of the $\frac{n}{2}$-th and $\left(\frac{n}{2}+1\right)$-th items, the two items on the middle.
The median $\tilde{x}$ in a particular sample $(x_1, x_2, \ldots, x_n)$, whose elements are arranged in ascending (or descending) order, is

$$\tilde{x} = \begin{cases} x_{m+1}, & \text{if } n = 2m+1, \\ \dfrac{x_{m+1}+x_m}{2}, & \text{if } n = 2m. \end{cases} \tag{16.123}$$

4. Range

$$R = \max_i X_i - \min_i X_i \qquad (i = 1, 2, \ldots, n). \tag{16.124a}$$

The range R of a particular sample $(x_1, x_2, \ldots, x_n)$ is

$$R = x_{\max} - x_{\min}. \tag{16.124b}$$

Every particular realization of a sample function is denoted by a lowercase letter, except the range R, i.e., for a particular sample $(x_1, x_2, \ldots, x_n)$ the particular values $\overline{x}$, s^2, $\tilde{x}$, and R are calculated.

i	x_i	i	x_i	i	x_i
1	1.01	6	1.00	11	1.00
2	1.02	7	0.99	12	1.00
3	1.00	8	1.01	13	1.02
4	0.98	9	1.01	14	1.00
5	0.99	10	1.00	15	1.01

■ Choosing a sample of 15 loudspeakers from a running production the interesting quantity X is the air gap induction B, measured in Tesla. From the measured data in the table to the left follows:
$\overline{x} = 1.0027$ or $\overline{x} = 1.0027$ with $x_0 = 1.00$;
$s^2 = 1.2095 \cdot 10^{-4}$ or $s^2 = 1.2076 \cdot 10^{-4}$ with $x_0 = 1.00$;
$\tilde{x} = 1.00$; $R = 0.04$.

16.3.2 Descriptive Statistics

16.3.2.1 Statistical Summarization and Analysis of Given Data

In order to describe statistically a property of a certain element this property must be characterized by a random variable X. Usually, the n measured or observed values x_i of the property X form the starting point of a statistical investigation, which is made to find some parameters of the distribution or the distribution itself of X.
Every measured sequence of size n can be considered as a random sample from an infinite population, if the experiment or the measurement could be repeated infinitely many times under the same conditions. Since the size n of a measuring sequence can be very large, the statistical investigation proceeds as follows:

1. Protocol, Prime Notation The measured or observed values x_i are recorded in a *protocol list.*

2. Intervals or Classes Grouping the measured n data x_i $(i = 1, 2, \ldots, n)$ of the sample into k subintervals, so-called *classes* or *class intervals* of equal length or width h; usually 10–20 classes.

3. Frequencies and Frequency Distribution The *absolute frequencies* h_j $(j = 1, 2, \ldots, k)$ are the numbers h_j of data (occupancy number) belonging to a given interval Δx_j. The ratios h_j/n (in %) are called *relative frequencies*. If the values h_j/n are represented over the classes as rectangles, then one gets a graphical representation of the given *frequency distribution*, and this representation is called a *histogram* (**Fig. 16.13a**). The values h_j/n can be considered as the empirical values of the probabilities or the density function $f(x)$.

Table 16.3 Frequency table

Class	h_i	h_i/n(%)	F_i (%)
50 – 70	1	0.8	0.8
71 – 90	1	0.8	1.6
91 – 110	2	1.6	3.2
111 – 130	9	7.2	10.4
131 – 150	15	12.0	22.4
151 – 170	22	17.6	40.0
171 – 190	30	24.0	64.0
191 – 210	27	21.6	85.6
211 – 230	9	7.2	92.8
231 – 250	6	4.8	97.6
251 – 270	3	2.4	100.0

4. Cumulative Frequency Adding the absolute or relative frequencies, gives the *cumulative absolute* or *relative frequency*

$$F_j = \frac{h_1 + h_2 + \cdots + h_j}{n}\,\% \quad (j = 1, 2, \ldots, k). \qquad (16.125)$$

A graphical representation of the empirical distribution function, which can be considered as an approximation of the unknown underlying distribution function $F(x)$ is shown in **Fig. 16.13b**.

■ Suppose during a study $n = 125$ measurements have been performed. The results spread in the interval from 50 to 270, so it is reasonable to divide this interval into $k = 11$ classes with a length $h = 20$. The *frequency table* see **Table 16.3**.

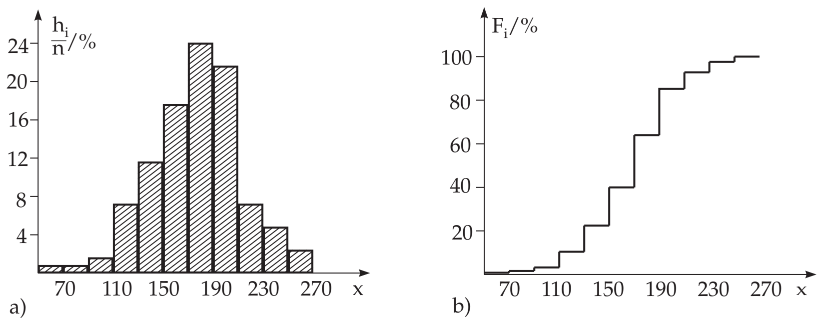


Figure 16.13

16.3.2.2 Statistical Parameters

After summarizing and analyzing the data of the sample as given in 16.3.2.1, p. 832, it follows the approximation of the parameters of the distribution belonging to the random variable by the following parameters:

1. Mean

Using all measured data from the sample directly, the sample mean is

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i. \qquad (16.126a)$$

Using the means $\overline{x}_j$ and frequencies h_j of the classes

$$\overline{x} = \frac{1}{n}\sum_{j=1}^{k} h_j \overline{x}_j. \qquad (16.126b)$$

2. Variance

Using all measured data directly the sample variance is

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2. \tag{16.127a}$$

Using the means $\overline{x}_j$ and frequencies h_j of the classes

$$s^2 = \frac{1}{n-1}\sum_{j=1}^{k} h_j(\overline{x}_j - \overline{x})^2. \tag{16.127b}$$

The *class midpoint* u_j (the midpoint of the corresponding interval) is also often used instead of $\overline{x}_j$.

3. Median

The median $\tilde{x}$ of a distribution is defined by

$$P(X < \tilde{x}) = \frac{1}{2}. \tag{16.128a}$$

The median may not be a uniquely determined point. The median of a sample is

$$\tilde{x} = \begin{cases} x_{m+1}, & \text{if } n = 2m+1, \\ \dfrac{x_{m+1} + x_m}{2}, & \text{if } n = 2m. \end{cases} \tag{16.128b}$$

4. Range

$$R = x_{\max} - x_{\min}. \tag{16.129}$$

5. Mode or Modal Value

is the data value that occurs with greatest frequency. It is denoted by D.

16.3.3 Important Tests

One of the fundamental problems of mathematical statistics is to draw conclusions about the population from the sample. There are two types of the most important questions:

1. The type of the distribution is known, and one wants to get some estimate for its parameters. A distribution can be characterized mostly quite well by the parameters μ and σ^2 (here μ is the exact value of the expected value, and σ^2 is the exact variance), consequently one of the most important questions is how good an estimation can be given for them, based on the samples.

2. Some hypotheses are known about these parameters, and it is desirable to check if they are true. The most often occurring questions are:

a) Is the expected value equal to a given number or not?

b) Are the expected values for two populations equal or not?

c) Does the distribution of the random variable with μ and σ^2 fit a given distribution or not? etc.

Because in observations and measurements, the normal distribution has a very important role, it is to be discussed the goodness of a fit test for a normal distribution. The basic idea can be used for other distributions, too.

16.3.3.1 Goodness of Fit Test for a Normal Distribution

There are different tests in mathematical statistics to decide if the data of a sample come from a normal distribution. Here a graphical one is discussed based on normal probability paper , and a numerical one based on the use of the chi-square distribution ("χ^2 test").

1. Goodness of Fit Test with Probability Paper

a) Principle of Probability Paper The x-axis in a right-angled coordinate system is scaled equidistantly, while the y-axis is scaled on the following way: It is divided equidistantly with respect to Z, but

scaled by

$$y = \Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{t^2}{2}}\, dt. \tag{16.130}$$

If a random variable X has a normal distribution with expected value μ and variance σ^2, then for its distribution function (see 16.2.4.2, p. 819)

$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right) = \Phi(z) \tag{16.131a}$$

holds, i.e.,

$$z = \frac{x-\mu}{\sigma} \tag{16.131b}$$

z	x
0	μ
1	$\mu+\sigma$
-1	$\mu-\sigma$

(16.131c)

must be valid, and so there is a linear relation between x and z and (16.131c).

b) Application of Probability Paper

Considering the data of the sample, calculating the cumulative relative frequencies according to (16.125), and sketch these onto the probability paper as the ordinates of the points with abscissae the upper class boundaries. If these points are approximately on a straight line, (with small deviations) then the random variable can be considered as a normal random variable (**Fig. 16.14**).

As it can be seen from **Fig. 16.14**, the distribution to which the data of **Table 16.3** belong, can be considered as a normal distribution. Furthermore one can see that $\mu \approx 176$, $\sigma \approx 37.5$ (from the x values belonging to the 0 and ± 1 values of Z).

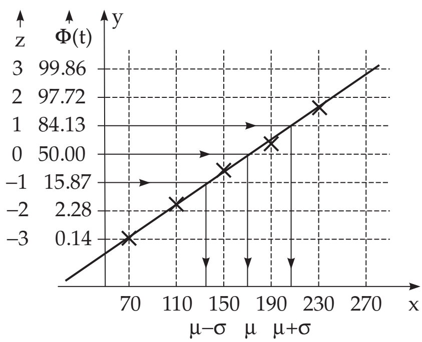


Figure 16.14

Remark: The values F_i of the relative cumulative frequencies can be plotted more easily on the probability paper, if its scaling is equidistant with respect to y, which means a non-equidistant scaling for the ordinates.

2. χ^2 test for Goodness of Fit

It is to check if a random variable X can be considered as normal in the sense of 16.2.4.2, p. 819. The range of X is divided into k classes and the upper limit of the j-th $(j = 1, 2, \ldots, k)$ class is denoted by ξ_j . Let p_j be the "theoretical" probability that X is in the j-th class, i.e.,

$$p_j = F(\xi_j) - F(\xi_{j-1}), \tag{16.132a}$$

where $F(x)$ is the distribution function of X $(j = 1, 2, \ldots, k\,;\ \xi_0$ is the lower limit of the first class with $F(\xi_0) = 0)$. Because X is supposed to be normal, then

$$F(\xi_j) = \Phi\left(\frac{\xi_j - \mu}{\sigma}\right) \tag{16.132b}$$

must hold, where $\Phi(x)$ is the distribution function of the standard normal distribution (see 16.2.4.2, p. 819). The parameters μ and σ^2 of the population are usually not known. Therefore $\overline{x}$ and s^2 are to be used as an approximation of them. The decomposition of the range of X should be made so that the expected frequencies for every class should exceed 5, i.e., if the size of the sample is n, then $np_j \geq 5$.

Now, considering the sample $(x_1, x_2, \ldots, x_n)$ of size n and calculating the corresponding frequencies h_j

(for the classes given above), then the random variable

$$\chi_S^2 = \sum_{j=1}^{k} \frac{(h_j - np_j)^2}{np_j} \tag{16.132c}$$

has approximately a χ^2 distribution with $m = k-1$ degrees of freedom, if μ and σ^2 are known, $m = k-2$ if one of them is estimated from the sample, and $m = k-3$ if both are estimated by $\overline{x}$ and s^2.
The test, if a random variable X has a normal distribution (χ^2-approximation test) consists in the comparison of the experimental χ_S^2 of the sample with the corresponding $\chi^2_{\alpha;m}$ from **Table 21.18**, p. 1135 for a given statistical *significance level* α and m degrees of freedom.
Now one decides a significant level α and one takes out of the **Table 21.18**, p. 1135 the quantile $\chi^2_{\alpha;k-i}$ of the corresponding χ^2 distribution (i depends on the number of unknown parameters). This means $P(\chi^2 \ge \chi^2_{\alpha;k-i}) = \alpha$ holds. Comparing the value χ_S^2 from (16.132c) and this quantile, and if

$$\chi_S^2 < \chi^2_{\alpha;k-i} \tag{16.132d}$$

holds, one can accept the assumption that the sample came from a normal distribution. This test is also called the χ^2 *test for goodness of fit.*

■ The following χ^2 test is based on the example on p. 833. The sample size is $n = 125$, with the mean $\overline{x} = 176.32$ and variance $s^2 = 36.70$. These values are used as approximations of the unknown parameters μ and σ^2 of the population. Determining the test statistic χ_S^2 according to (16.132c) after performing the calculations according to (16.132a) and (16.132b), yields the data in **Table 16.4**.

Table 16.4 χ^2 test

ξ_j	h_j	$\frac{\xi_j-\mu}{\sigma}$	$\Phi\left(\frac{\xi_j-\mu}{\sigma}\right)$	p_j	np_j	$\frac{(h_j-np_j)^2}{np_j}$
70	1 (Σ = 13)	−2.90	0.0019	0.0019	0.2375 (Σ = 12.9750)	0.00005
90	1	−2.35	0.0094	0.0075	0.9375	
110	2	−1.81	0.0351	0.0257	3.2125	
130	9	−1.26	0.1038	0.0687	8.5857	
150	15	−0.72	0.2358	0.1320	16.5000	0.1635
170	22	−0.17	0.4325	0.1967	24.5875	0.2723
190	30	0.37	0.6443	0.2118	26.4750	0.4693
210	27	0.92	0.8212	0.1769	22.1125	1.0803
230	9	1.46	0.9279	0.1067	13.3375	1.4106
250	6 (Σ = 9)	2.01	0.9778	0.0499	6.2375 (Σ = 8.3375)	0.0526
270	3	2.55	0.9946	0.0168	2.1000	
						$\chi_S^2 = 3.4486$

It follows from the last column that $\chi_S^2 = 3.4486$. Because of the requirement $np_j \ge 5$, the number of classes is reduced from $k = 11$ to $k^* = k - 4 = 7$. Since for the calculation of the theoretical frequencies np_j the estimated values $\overline{x}$ and s^2 of the sample are used instead of μ and σ^2 of the population, so the number of degrees of freedom of the corresponding χ^2 distribution is reduced by 2. The critical value is the quantile $\chi^2_{\alpha;k^*-1-2}$. For $\alpha = 0.05$ one gets $\chi^2_{0.05;4} = 9.5$ from **Table 21.18**, p. 1135, so because of the inequality $\chi_S^2 < \chi^2_{0.05;4}$ there is no contradiction to the assumption that the sample is from a population with a normal distribution.

16.3.3.2 Distribution of the Sample Mean

Let X be a continuous random variable. Suppose arbitrarily many samples of size n can be taken from the corresponding population. Then the sample mean is also a random variable $\overline{X}$, and it is also continuous.

1. Confidence Probability of the Sample Mean

If X has a normal distribution with parameters μ and σ^2, then $\overline{X}$ is also a normal random variable with parameters μ and σ^2/n, i.e., the density function $\overline{f}(x)$ of $\overline{X}$ is concentrated more around μ than the density function $f(x)$ of the population. For any value $\varepsilon > 0$

$$P(|X-\mu| \le \varepsilon) = 2\Phi\left(\frac{\varepsilon}{\sigma}\right) - 1, \quad P(|\overline{X}-\mu| \le \varepsilon) = 2\Phi\left(\frac{\varepsilon\sqrt{n}}{\sigma}\right) - 1 \tag{16.133}$$

holds. Consequently with increasing sample size n the probability that the sample mean is a good approximation of μ is also increasing.

■ For $\varepsilon = \frac{1}{2}\sigma$ from (16.133) $P\left(|\overline{X}-\mu| \le \frac{1}{2}\sigma\right) = 2\Phi\left(\frac{1}{2}\sqrt{n}\right) - 1$, and for different values of n one gets the values listed in **Table 16.5**. As can be seen from **Table 16.5**, e.g., that with a sample size $n = 49$, the probability that the sample mean $\overline{x}$ differs from μ by less than $\pm\frac{1}{2}\sigma$ is 99.95 %.

Table 16.5 Confidence level for the sample mean

n	$P\left(\lvert\overline{X}-\mu\rvert \le \frac{1}{2}\sigma\right)$
1	38.29 %
4	68.27 %
16	95.45 %
25	98.76 %
49	99.96 %

2. Sample Mean Distribution for Arbitrary Distribution of the Population

The random variable $\overline{X}$ has an approximately normal distribution with parameters μ and σ^2/n for any distribution of the population with expected value μ and variance σ^2. This fact is based on the central limit theorem.

16.3.3.3 Confidence Limits for the Mean

1. Confidence Interval for the Mean with a Known Variance σ^2

If X is a random variable with parameters μ and σ^2, then according to 16.3.3.2, p. 836, $\overline{X}$ is approximately a normal random variable with parameters μ and σ^2/n. Then substitution of

$$\overline{Z} = \frac{\overline{X}-\mu}{\sigma}\sqrt{n} \tag{16.134}$$

yields a random variable $\overline{Z}$ which has approximately a standard normal distribution, therefore

$$P(|\overline{Z}| \le \varepsilon) = \int_{-\varepsilon}^{\varepsilon} \varphi(x)\,dx = 2\Phi(\varepsilon) - 1. \tag{16.135}$$

If the given significance level is α, namely,

$$P(|\overline{Z}| \le \varepsilon) = 1 - \alpha, \tag{16.136}$$

is required, then $\varepsilon = \varepsilon(\alpha)$ can be determined from (16.135), e.g., from **Table 21.17**, p. 1133, for the standard normal distribution. From $|\overline{Z}| \le \varepsilon(\alpha)$ and from (16.134) the relation

$$\mu = \overline{x} \pm \frac{\sigma}{\sqrt{n}}\varepsilon(\alpha) \tag{16.137}$$

follows. The values $\overline{x} \pm \frac{\sigma}{\sqrt{n}}\varepsilon(\alpha)$ in (16.137) are called *confidence limits for the expected value* and the interval between them is called a *confidence interval for the expected value* μ with a known σ^2 and given significance level α. In other words: The expected value μ is between the confidence limits (16.137) with a probability $1 - \alpha$.

Remark: If the sample size is large enough, then s^2 can be used instead of σ^2 in (16.137). The sample size is considered to be large, if $n > 100$, but in practice, depending on the actual problem, it is considered to be sufficiently large if $n > 30$. If n is not large enough, then in the case of a normally distributed population the t distribution is applied to determine the confidence limits as in (16.140).

2. Confidence Interval for the Expected Value with an Unknown Variance σ^2

If the population is approximately normally distributed and its variance σ^2 is unknown, then it is replaced by the sample variance s^2 and instead of (16.134) one gets the random variable

$$T = \frac{\overline{X} - \mu}{s}\sqrt{n}\,, \tag{16.138}$$

which has a t distribution (see 16.2.4.8, p. 824) with $m = n - 1$ degrees of freedom. Here n is the size of the sample. If n is large, e.g., $n > 100$ holds, then T can be considered as a normal random variable as Z in (16.134). For a given significance level α holds

$$P(\,|T| \le \varepsilon) = \int\limits_{-\varepsilon}^{\varepsilon} f_t(x)\,dx = P\left(\frac{|\overline{X} - \mu|}{s}\sqrt{n} \le \varepsilon\right) = 1 - \alpha. \tag{16.139}$$

From (16.139) follows that $\varepsilon = \varepsilon(\alpha, n) = t_{\alpha/2;n-1}$, where $t_{\alpha/2;n-1}$ is the quantile of the t distribution (with $n-1$ degrees of freedom) for the significance level α (**Table 21.20**, p. 1138). From $|T| = t_{\alpha/2;n-1}$

$$\mu = \overline{x} \pm \frac{s}{\sqrt{n}} t_{\alpha/2;n-1} \tag{16.140}$$

follows. The values $\overline{x} \pm \frac{s}{\sqrt{n}} t_{\alpha/2;n-1}$ are called the *confidence limits for the expected value* μ of the distribution of the population with an unknown variance σ^2 and with a given significance level α. The interval between these limits is the *confidence interval*.

■ A sample contains the following 6 measured values: 0.842; 0.846; 0.835; 0.839; 0.843; 0.838. Then $\overline{x} = 0.8405$ and $s = 0.00394$ is obtained.
What is the maximum deviation of the sample mean $\overline{x}$ from the expected value μ of the population distribution, if the significance level α is given as 5 % or 1 % ?
1. $\alpha = 0.05$: one reads from **Table 21.20**, p. 1138, that $t_{\alpha/2;5} = 2.57$, consequently
$|\overline{X} - \mu| \le 2.57 \cdot 0.00394/\sqrt{6} = 0.0042$. Thus, the sample mean $\overline{x} = 0.8405$ differs from the expected value μ by less than ± 0.0042 with a probability 95 %.
2. $\alpha = 0.01$: $t_{\alpha/2;5} = 4.03$; $|\overline{X} - \mu| \le 4.03 \cdot 0.00394/\sqrt{6} = 0.0065$, i.e., the sample mean $\overline{x}$ differs from μ by less than ± 0.0065 with a probability 99 %.

16.3.3.4 Confidence Interval for the Variance

If the random variable X has a normal distribution with parameters μ and σ^2, then the random variable

$$\chi^2 = (n-1)\frac{s^2}{\sigma^2} \tag{16.141}$$

has a χ^2 distribution with $m = n - 1$ degrees of freedom, where n is the sample size and s^2 is the sample deviation. $f_{\chi^2}(x)$ denotes the density function of the χ^2 distribution in **Fig. 16.15**, and one sees that

$$P(\chi^2 < \chi^2_u) = P(\chi^2 > \chi^2_o) = \frac{\alpha}{2}. \tag{16.142}$$

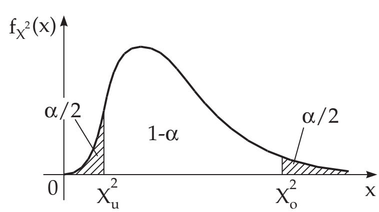


Figure 16.15

Thus, using the quantiles of the χ^2 distribution (see **Table 21.18**, p. 1135) gives

$$\chi^2_u = \chi^2_{1-\alpha/2;n-1}, \quad \chi^2_o = \chi^2_{\alpha/2;n-1}. \tag{16.143}$$

Considering (16.141) one gets the following estimation for the unknown variance σ^2 of the population

distribution with a significance level α:

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2;n-1}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{1-\alpha/2;n-1}}. \tag{16.144}$$

The confidence interval given in (16.144) for σ^2 is fairly large for small sample sizes.

■ For the numerical example (p. 838) with 6 measured values and for $\alpha = 5\%$ one gets from **Table 21.18**, p. 1135, $\chi^2_{0.025;5} = 0.831$ and $\chi^2_{0.975;5} = 12.8$, so it follows from (16.144) that $0.625 \cdot s \le \sigma \le 2.453 \cdot s$ with $s = 0.00394$.

16.3.3.5 Structure of Hypothesis Testing

A statistical hypothesis testing has the following structure:

1. First a hypothesis H is to be developed that the sample belongs to a population with some given properties, e.g.,
H: The population distribution has a normal distribution with parameters μ and σ^2 (or another given distribution), or
H: For the non-known μ an approximative value (estimate value) μ_0 is inserted which can be got ,e.g., by rounding off the sample mean $\overline{x}$, or
H: Two populations have the same expected value, $\mu_1 - \mu_2 = 0$, etc.

2. A confidence interval B is to be defined, based on the hypothesis H (in general with tables). The value of the sample function should be in this interval with the given probability, e.g., with probability 99% for $\alpha = 0.01$.

3. Calculating the value of the sample function and accepting the hypothesis if this value is in the given interval B, otherwise rejecting it.

■ Test of the hypothesis H: $\mu = \mu_0$ with a significance level α.

The random variable $T = \dfrac{\overline{X} - \mu_0}{s}\sqrt{n}$ has a t distribution with $m = n-1$ degrees of freedom according to 16.3.3.3, p. 837. It follows from this that this hypothesis is to reject, if $\overline{x}$ is not in the confidence interval given by (16.140), i.e., if

$$|\overline{x} - \mu_0| \ge \frac{s}{\sqrt{n}}\, t_{\alpha/2;n-1} \tag{16.145}$$

holds. One says that there is a *significant difference.* For further problems about tests see [16.16].

16.3.4 Correlation and Regression

Correlation analysis is used to determine some kind of dependence between two or more properties of the population from the experimental data. The form of this dependence between these properties is determined with the help of *regression analysis.*

16.3.4.1 Linear Correlation of two Measurable Characters

1. Two-Dimensional Random Variable

In the following the formulas are mostly used for continuous random variables, but it is easy to replace them by the corresponding formulas for discrete variables. Suppose that X and Y, as a two-dimensional random variable (X, Y), have the joint distribution function

$$F(x,y) = P(X \le x, Y \le y) = \int\limits_{-\infty}^{x}\int\limits_{-\infty}^{y} f(x,y)\,dx\,dy, \tag{16.146a}$$

$$F_1(x) = P(X \le x, Y < \infty), \quad F_2(y) = P(X < \infty, Y \le y). \tag{16.146b}$$

The random variables X and Y are called *independent of each other* if

$$F(x,y) = F_1(x) \cdot F_2(y) \tag{16.147}$$

holds. The fundamental quantities assigned to X and Y determined from their joint density function $f(x, y)$ are:

1. Expected Values

$$\mu_X = E(X) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x\, f(x,y)\,dx\,dy, \quad (16.148a) \qquad \mu_Y = E(Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y\, f(x,y)\,dx\,dy, \quad (16.148b)$$

2. Variances

$$\sigma_X^2 = E((X-\mu_X)^2), \quad (16.149a) \qquad \sigma_Y^2 = E((Y-\mu_Y)^2). \quad (16.149b)$$

3. Covariance

$$\sigma_{XY} = E\left((X-\mu_X)(Y-\mu_Y)\right). \quad (16.150)$$

4. Correlation Coefficient

$$\varrho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}. \quad (16.151)$$

It is assumed that every expected value above exists. The covariance can also be calculated by the formula

$$\sigma_{XY} = E(XY) - \mu_X\mu_Y \text{ where } E(XY) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\, f(x,y)\,dx\,dy. \quad (16.152)$$

The correlation coefficient is a measure of the linear dependence of X and Y, because of the following facts:
All points (X, Y) are on one line with probability 1 if $\varrho_{XY}^2 = 1$ holds. If X and Y are independent random variables, then their covariance is equal to zero, $\varrho_{XY} = 0$. From $\varrho_{XY} = 0$, it does not follow that X and Y are independent, but it does if they have a *two-dimensional normal distribution* which is defined by the density function

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\varrho_{XY}^2}} \exp\left[-\frac{1}{2(1-\varrho_{XY}^2)}\left(\frac{(x-\mu_X)^2}{\sigma_X^2} - 2\frac{\varrho_{XY}(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right)\right]. \quad (16.153)$$

2. Test for Independence of two Variable

In practical problems often the question arises of whether the variables X and Y can be considered independent with $\varrho_{XY} = 0$, if the sample with size n and with the measured values (x_i, y_i) $(i = 1, 2, \ldots, n)$ comes from a two-dimensional normal distributed population. The test is performed in the following way:

1. Setting up the hypothesis H: $\varrho_{XY} = 0$.

2. Determining a significance level α and determining the quantile $t_{\alpha,m}$ of the t distribution from **Table 21.20**, p. 1138, for $m = n - 2$.

3. Calculation of the *empirical correlation coefficients* r_{xy} and calculation of the test statistics (sample function)

$$t = \frac{r_{xy}\sqrt{n-2}}{\sqrt{1-r_{xy}^2}} \quad (16.154a) \qquad \text{with} \quad r_{xy} = \frac{\sum\limits_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2 \sum\limits_{i=1}^{n}(y_i-\overline{y})^2}}. \quad (16.154b)$$

4. Rejection of the hypothesis if $|t| \ge t_{\alpha,m}$ holds.

16.3.4.2 Linear Regression for two Measurable Characters

1. Determination of the Regression Line

If a certain dependence is detected between the variables X and Y by the correlation coefficient, then the next problem is to find the functional dependence $Y = f(X)$. Here mostly linear dependence is considered.

In the simplest case of *linear regression* it is supposed that for any fixed value of x the random variable Y in the population has a normal distribution with the expected value

$$E(Y) = a + bx \qquad (16.155)$$

and a variance σ^2 independent of x. The relation (16.155) means that the mean value of the random variable Y depends linearly on the fixed value of x. The values of the parameters a, b and σ^2 of the population are usually unknown. To estimate them approximately from a sample with values (x_i, y_i) $(i = 1, 2, \ldots, n)$ the *least squares method* can be used. The least squares method requires that

$$\sum_{i=1}^{n}[y_i - (a + bx_i)]^2 = \min! \qquad (16.156)$$

is valid so one gets the estimates

$$\tilde{b} = \frac{\sum\limits_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum\limits_{i=1}^{n}(x_i - \overline{x})^2}, \quad \tilde{a} = \overline{y} - \tilde{b}\overline{x}, \quad \tilde{\sigma}^2 = \frac{n-1}{n-2}s_y^2(1 - r_{xy}^2) \quad \text{with} \qquad (16.157a)$$

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \quad s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2. \qquad (16.157b)$$

The empirical correlation coefficient r_{xy} is given in (16.154b). The coefficients $\tilde{a}$ and $\tilde{b}$ are called *regression coefficients*. The line $y(x) = \tilde{a} + \tilde{b}x$ is called the *regression line*.

2. Confidence Intervals for the Regression Coefficients

The next question is, after the determination of the regression coefficients $\tilde{a}$ and $\tilde{b}$, how well do the estimates approximate the theoretical values a and b. Therefore the test variables are to be formed as

$$t_b = (\tilde{b} - b)\frac{s_x\sqrt{n-2}}{s_y\sqrt{1 - r_{xy}^2}} \quad (16.158a) \qquad \text{and} \qquad t_a = (\tilde{a} - a)\frac{s_x\sqrt{n-2}}{s_y\sqrt{1 - r_{xy}^2}}\,\frac{\sqrt{n}}{\sqrt{\sum\limits_{i=1}^{n}{x_i}^2}}. \qquad (16.158b)$$

These are realizations of random variables having a t distribution with $m = n - 2$ degrees of freedom. For a given significance level α the quantile $t_{\alpha/2;m}$ is taken from **Table 21.20**, p. 1138, and because $P(\,|t| < t_{\alpha/2;m}) = 1 - \alpha$ holds for $t = t_a$ and $t = t_b$:

$$|\tilde{b} - b| < t_{\alpha/2;n-2}\,\frac{s_y\sqrt{1 - r_{xy}^2}}{s_x\sqrt{n-2}}, \quad (16.159a) \quad |\tilde{a} - a| < t_{\alpha/2;n-2}\,\frac{s_y\sqrt{1 - r_{xy}^2}\cdot\sqrt{\sum\limits_{i=1}^{n}{x_i}^2}}{s_x\sqrt{n-2}\cdot\sqrt{n}}. \qquad (16.159b)$$

A *confidence region* for the regression line $y = a + bx$ can be determined with the confidence interval given in (16.159a,b) for a and b (see Lit. [16.4], [16.11]).

16.3.4.3 Multidimensional Regression

1. Functional Dependence

Suppose that there is a functional dependence between the characters $X_1, X_2, \dots, X_n$, and Y, which is described by the theoretical regression function

$$y = f(x_1, x_2, \dots, x_n) = \sum_{j=0}^{s} a_j g_j(x_1, x_2, \dots, x_n). \tag{16.160}$$

The functions $g_j(x_1, x_2, \dots, x_n)$ are known functions of n independent variables. The coefficients a_j in (16.160) are constant multipliers in this linear combination. The expression (16.160) is also called *linear regression*, although the functions g_j can be arbitrary.

■ The function $f(x_1, x_2) = a_0 + a_1x_1 + a_2x_2 + a_3{x_1}^2 + a_4{x_2}^2 + a_5x_1x_2$, which is a complete quadratic polynomial of two variables with $g_0 = 1$, $g_1 = x_1$, $g_2 = x_2$, $g_3 = {x_1}^2$, $g_4 = {x_2}^2$, and $g_5 = x_1x_2$, is an example for a theoretical linear regression function.

2. Writing in Vector Form

It is useful to write formulas in vector form in the multidimensional case

$$\underline{\mathbf{x}} = (x_1, x_2, \dots, x_n)^{\mathrm{T}}, \tag{16.161}$$

so, (16.160) now has the form:

$$y = f(\underline{\mathbf{x}}) = \sum_{j=0}^{s} a_j g_j(\underline{\mathbf{x}}). \tag{16.162}$$

3. Solution and Normal Equation System

The theoretical dependence (16.160) cannot be determined by the measured values

$$(\underline{\mathbf{x}}^{(i)}, f_i), \quad (i = 1, 2, \dots, N) \tag{16.163a}$$

because of random measuring errors. Looking for the solution in the form

$$y = \tilde{f}(\underline{\mathbf{x}}) = \sum_{j=0}^{s} \tilde{a}_j g_j(\underline{\mathbf{x}}) \tag{16.163b}$$

and using the least squares method (see 16.3.4.2, **1.**, p. 841) the coefficients $\tilde{a}_j$ as the estimations of the theoretical coefficients a_j can be determined, from the equation

$$\sum_{i=1}^{N} \left[f_i - \tilde{f}\left(\underline{\mathbf{x}}^{(i)}\right)\right]^2 = \min!\,. \tag{16.163c}$$

Introducing the notation

$$\underline{\tilde{\mathbf{a}}} = \begin{pmatrix} \tilde{a}_0 \\ \tilde{a}_1 \\ \vdots \\ \tilde{a}_s \end{pmatrix}, \quad \underline{\mathbf{f}} = \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_N \end{pmatrix}, \quad \mathbf{G} = \begin{pmatrix} g_0\left(\underline{\mathbf{x}}^{(1)}\right) & g_1\left(\underline{\mathbf{x}}^{(1)}\right) & \dots & g_s\left(\underline{\mathbf{x}}^{(1)}\right) \\ g_0\left(\underline{\mathbf{x}}^{(2)}\right) & g_1\left(\underline{\mathbf{x}}^{(2)}\right) & \dots & g_s\left(\underline{\mathbf{x}}^{(2)}\right) \\ \vdots & \vdots & \ddots & \vdots \\ g_0\left(\underline{\mathbf{x}}^{(N)}\right) & g_1\left(\underline{\mathbf{x}}^{(N)}\right) & \dots & g_s\left(\underline{\mathbf{x}}^{(N)}\right) \end{pmatrix} \tag{16.163d}$$

one gets from (16.163c) the so-called *normal system of equations*

$$\mathbf{G}^{\mathbf{T}}\mathbf{G}\underline{\tilde{\mathbf{a}}} = \mathbf{G}^{\mathbf{T}}\underline{\mathbf{f}} \tag{16.163e}$$

to determine $\underline{\tilde{\mathbf{a}}}$. The matrix $\mathbf{G}^{\mathbf{T}}\mathbf{G}$ is symmetric, so the Cholesky method (see 19.2.1.2, p. 958) is especially good to solve (16.163e).

■ Consider the sample whose result is given in the next table. Determine the coefficients of the regression function (16.164):

x_1	5	3	5	3
x_2	0.5	0.5	0.3	0.3
$f(x_1,x_2)$	1.5	3.5	6.2	3.2

$$\tilde{f}(x_1, x_2) = a_0 + a_1 x_1 + a_2 x_2. \qquad (16.164)$$

From (16.163d) it follows that

$$\underline{\tilde{\mathbf{a}}} = \begin{pmatrix} \tilde{a}_0 \\ \tilde{a}_1 \\ \tilde{a}_2 \end{pmatrix}, \quad \underline{\mathbf{f}} = \begin{pmatrix} 1.5 \\ 3.5 \\ 6.2 \\ 3.2 \end{pmatrix}, \quad \mathbf{G} = \begin{pmatrix} 1 & 5 & 0.5 \\ 1 & 3 & 0.5 \\ 1 & 5 & 0.3 \\ 1 & 3 & 0.3 \end{pmatrix} \qquad (16.165)$$

and (16.163e) is

$$\begin{aligned} 4\tilde{a}_0 + 16\,\tilde{a}_1 + 1.6\,\tilde{a}_2 &= 14.4, \\ 16\tilde{a}_0 + 68\,\tilde{a}_1 + 6.4\,\tilde{a}_2 &= 58.6, \\ 1.6\tilde{a}_0 + 6.4\tilde{a}_1 + 0.68\tilde{a}_2 &= 5.32, \end{aligned} \quad \text{i.e.,} \quad \begin{aligned} \tilde{a}_0 &= 7.0, \\ \tilde{a}_1 &= 0.25, \\ \tilde{a}_2 &= -11. \end{aligned} \qquad (16.166)$$

4. Remarks

1. To determine the regression coefficients one starts with the interpolation $\tilde{f}\left(\underline{\mathbf{x}}^{(i)}\right) = f_i$ $(i = 1, 2, \ldots, N)$, i.e., with

$$\mathbf{G}\underline{\tilde{\mathbf{a}}} = \underline{\mathbf{f}}. \qquad (16.167)$$

In the case $s < N$, (16.167) is an over determined system of equations which can be solved by the Householder method (see 19.6.2.2, p. 985). The multiplication of (16.167) by $\mathbf{G}^{\mathbf{T}}$ to get (16.163e) is also called *Gauss transformation*. If the columns of the matrix $\mathbf{G}$ are linearly independent, i.e., rank $\mathbf{G} = s + 1$ holds, then the normal system of equations (16.163e) has a unique solution, which coincides with the result of (16.167) got by the Householder method.

2. Also in the multidimensional case, it is possible to determine confidence intervals for the regression coefficients with the t distribution, analogously to (16.159a,b).

3. By the help of the F distribution (see 16.2.4.7, p. 823), it is possible to use a so-called *equivalence test* to analyze the assumption (16.163b). This test shows if a solution in the form (16.163b) but with less terms yields a sufficient approximation to the theoretic regression function (16.160). (see Lit. [16.10]).

16.3.5 Monte Carlo Methods

16.3.5.1 Simulation

Simulation methods are based on constructing equivalent mathematical models. These models are then easily analyzed by computer. In such cases, one talks about *digital simulation*. A special case is given by *Monte Carlo methods* when certain quantities of the model are randomly selected. These random elements are selected by using *random numbers*.

16.3.5.2 Random Numbers

Random numbers are realizations of certain random quantities (see 16.2.2, p. 811) satisfying given distributions. In this way it is possible to distinguish different kinds of random numbers.

1. Uniformly Distributed Random Numbers

These numbers are uniformly distributed in the interval $[0, 1]$, they are realizations of the random variable X with the following density function $f_0(x)$ and distribution function $F_0(x)$:

$$f_0(x) = \begin{cases} 1 & \text{for} \quad 0 \le x \le 1, \\ 0 & \text{otherwise;} \end{cases} \qquad F_0(x) = \begin{cases} 0 & \text{for} \quad 0 \le x, \\ x & \text{for} \quad 0 < x \le 1, \\ 1 & \text{for} \quad x \ge 1. \end{cases} \qquad (16.168)$$

1. Method of the Inner Digits of Squares A simple method to generate random numbers is suggested by J. v. Neumann. It is also called the *method of the inner digits of squares*, and it starts from a decimal number $z \in (0, 1)$ which has $2n$ digits. First z^2 is formed, getting a decimal number

which has $4n$ digits. Erasing its first and its last n digits, so one again has a number with $2n$ digits. To get further numbers, this procedure is repeated. In this way $2n$ digit decimal numbers are produced from the interval $[0, 1]$ which can be considered as random numbers with a uniform distribution. The value of $2n$ is selected according to the largest number representable in the computer. For example, choosing $2n = 10$. This procedure is seldom recommended, since it produces more smaller numbers than it should. Several other different methods have been developed.

■ $2n = 4$:
$z = z_0 = 0,1234$, $z_0^2 = 0,01\boxed{5227}56$,
$z = z_1 = 0,5227$, $z_1^2 = 0,27\boxed{3215}29$,
$z = z_2 = 0,3215$ etc.
The first three random numbers are z_0, z_1 and z_2.

2. Congruence Method The so-called *congruence method* is widely used: A sequence of integers z_i $(i = 0, 1, 2, \ldots)$ is formed by the recursion formula

$$z_{i+1} \equiv c \cdot z_i \mod m. \tag{16.169}$$

Here z_0 is an arbitrary positive number and c and m denote positive integers, which must be suitably chosen. For z_{i+1} one takes the smallest non-negative integer satisfying the congruence (16.169). The numbers z_i/m are between 0 and 1 and can be used for uniformly distributed random numbers.

3. Remarks

a) Choosing $m = 2^r$, where r is the number of bits in a computer word, e.g., $r = 40$. Then the number c is to be chosen in the order of $\sqrt{m}$.

b) Random number generators using certain algorithms produce so-called *pseudo-random numbers*.

c) On calculators and also in computers, "ran" or "rand" is used for generating random numbers.

2. Random Numbers with other Distributions

To get random numbers with an arbitrary distribution function $F(x)$ the following procedure is in use: Consider a sequence of uniformly distributed random numbers $\xi_1, \xi_2, \ldots$ from $[0, 1]$. With these numbers the numbers $\eta_i = F^{-1}(\xi_i)$ are formed for $i = 1, 2, \ldots$. Here $F^{-1}(x)$ is the inverse function of the distribution function $F(x)$. Then one gets:

$$P(\eta_i \le x) = P(F^{-1}(\xi_i) \le x) = P(\xi_i \le F(x)) = \int_0^{F(x)} f_0(t)\,dt = F(x), \tag{16.170}$$

i.e., the random numbers $\eta_1, \eta_2, \ldots$ satisfy a distribution with the distribution function $F(x)$.

3. Tables and Application of Random Numbers

1. Construction Random number tables can be constructed in the following way. Ten identical chips are indexed by the numbers $0, 1, 2, \ldots, 9$. Placing them into a box and shuffle them. One of them is then selected, and its index is written into the table. Then the chip is to be replaced into the box, again shuffling, and choosing the next one. In this way a sequence of random numbers is produced, which is written in groups (for easier usage) into the table. In **Table 21.21**, p. 1139, four random digits form a group.

In the procedure, it must be guaranteed that the digits $0, 1, 2, \ldots, 9$ always have equal probability.

2. Application of Random Numbers The use of a table of random numbers is demonstrated with an example.

■ Suppose one chooses randomly $n = 20$ items from a population of $N = 250$ items. The objects are renumbered from 000 to 249. Then a number is chosen in an arbitrary column or row in **Table 21.21**, p. 1139, and a rule is determined of how the remaining 19 random numbers should be chosen, e.g., vertically, horizontally or diagonally. Only the first three digits are considered from these random

numbers, and they are used only if they form a number smaller than 250.

16.3.5.3 Example of a Monte Carlo Simulation

The approximate evaluation of the integral

$$I = \int_0^1 g(x)\,dx \tag{16.171}$$

is considered as an example of the use of uniformly distributed random numbers in a simulation. Two solution methods are discussed here.

1. Applying the Relative Frequency

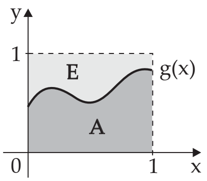


Figure 16.16

Supposing $0 \le g(x) \le 1$ holds; this condition can always be guaranteed by a transformation (see (16.175), p. 845). Then the integral I is an area inside the unit square E **(Fig. 16.16)**. Considering the numbers of a sequence of uniformly distributed random numbers from the interval $[0, 1]$ in pairs as the coordinates of points of the unit square E, one gets n points P_i $(i = 1, 2, \ldots, n)$. Denoting by m the number of points inside the area A, then considering the notion of the relative frequency (see 16.2.1.2, p. 808) for the integral follows:

$$\int_0^1 g(x)\,dx \approx \frac{m}{n}. \tag{16.172}$$

To achieve relatively good accuracy with the ratio in (16.172), a very large number of random numbers is necessary. This is the reason why one is looking for possibilities to improve the accuracy. One of these methods is the following Monte Carlo method. Some others can be found in the literature (see Lit. [16.20]).

2. Approximation by the Mean Value

To determine (16.171), one starts with n uniformly distributed random numbers $\xi_1, \xi_2, \ldots, \xi_n$ as the realization of the uniformly distributed random variable X. Then the values $g_i = g(\xi_i)$ $(i = 1, 2, \ldots, n)$ are realizations of the random variable $g(X)$, whose expectation according to formula (16.49a,b), p. 813, is:

$$E(g(X)) = \int_{-\infty}^{\infty} g(x) f_0(x) dx = \int_0^1 g(x) dx \approx \frac{1}{n} \sum_{i=1}^{n} g_i. \tag{16.173}$$

This method, which uses a sample to obtain the mean value, is also called the *usual Monte Carlo method.*

16.3.5.4 Application of the Monte Carlo Method in Numerical Mathematics

1. Evaluation of Multiple Integrals

First, it is to show how to transform a definite integral of one variable (16.174a) into an expression which contains the integral (16.174b).

$$I^* = \int_a^b h(x)\,dx \tag{16.174a}$$

$$I = \int_0^1 g(x)\,dx \quad \text{with} \quad 0 \le g(x) \le 1. \tag{16.174b}$$

Then the Monte Carlo method given in 16.3.5.3 can be applied introducing the following notation:

$$x = a + (b - a)u, \qquad m = \min_{x \in [a,b]} h(x), \qquad M = \max_{x \in [a,b]} h(x). \tag{16.175}$$

Then (16.174a) becomes

$$I^* = (M-m)(b-a)\int\limits_0^1 \frac{h(a+(b-a)u)-m}{M-m}\,du + (b-a)m, \tag{16.176}$$

where the integrand $\dfrac{h(a+(b-a)u)-m}{M-m} = g(u)$ satisfies the relation $0 \le g(u) \le 1$.

■ The approximate evaluation of multiple integrals with Monte Carlo methods is demonstrated by an example of a double integral

$$V = \iint\limits_S h(x,y)\,dx\,dy \quad \text{with} \quad h(x,y) \ge 0. \tag{16.177}$$

S denotes a plane surface domain given by the inequalities $a \le x \le b$ and $\varphi_1(x) \le y \le \varphi_2(x)$, where $\varphi_1(x)$ and $\varphi_2(x)$ denote given functions. Then V can be considered as the volume of a cylindrical solid K, which stands perpendicular to the x, y plane and its upper surface is given by $h(x,y)$. If $h(x,y) \le e$ holds, then this solid is in a block Q given by the inequalities $a \le x \le b,\ c \le y \le d,\ 0 \le z \le e \quad (a, b, c, d, e \text{ const})$. After a transformation similar to (16.175), one gets from (16.177) an expression containing the integral

$$V^* = \iint\limits_{S^*} g(u,v)\,du\,dv \quad \text{with} \quad 0 \le g(u,v) \le 1, \tag{16.178}$$

where V^* can be considered as the volume of a solid K^* in the three-dimensional unit cube. The integral (16.178) is approximated by the Monte Carlo method in the following way:
Triplets of the numbers of a sequence of uniformly distributed random numbers from the interval $[0,1]$ are considered as the coordinates of points $P_i \quad (i = 1, 2, \ldots, n)$ of the unit cube, and count how many points among P_i belong to the solid K^*. If m point belong to K^*, then analogously to (16.172)

$$V^* \approx \frac{m}{n}\,. \tag{16.179}$$

Remark: In definite integrals with one integration variable one should use the methods given in 19.3, p. 963. For the evaluation of multiple integrals, the Monte Carlo method is still often recommended.

2. Solution of Partial Differential Equations with the Random Walk Process

The Monte Carlo method can be used for the approximate solution of partial differential equations with the *random walk process*.

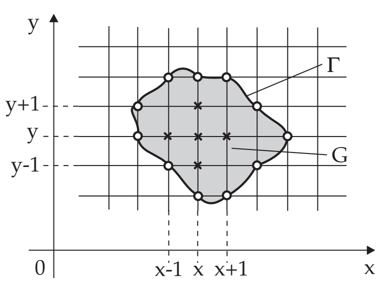


Figure 16.17

a) Example of a Boundary Value Problem: Consider the following boundary value problem as an example:

$$\Delta u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 \quad \text{for} \quad (x,y) \in G, \tag{16.180a}$$

$$u(x,y) = f(x,y) \quad \text{for} \quad (x,y) \in \Gamma. \tag{16.180b}$$

Here G is a simply connected domain in the x, y plane; Γ denotes the boundary of G **(Fig.16.17)**. Similarly to the difference method in paragraph 19.5.1, p. 976 G is covered by a quadratic lattice, where it can be assumed, without loss of generality, that the step size can be chosen as $h = 1$.

In this way interior lattice points $P(x,y)$ and boundary points R_i are obtained. The boundary points R_i, which are at the same time also lattice points, are considered in the following discussion as points of the boundary Γ of G, i.e.:

$$u(R_i) = f(R_i) \quad (i = 1, 2, \ldots, N) \tag{16.181}$$

b) Solution Principle: One imagines that a particle starts a *random walk* from an interior point $P(x, y)$. That is:

1. The particle moves randomly from $P(x, y)$ to one of the four neighboring points. To each of these four grid points 1/4 is assigned as the probability to move into them.

2. If the particle reaches a boundary point R_i, then the random walk terminates there with probability one.

It can be proven that a particle starting at any interior point P reaches a boundary point R_i after a finite number of steps with probability one. Denoting by

$$p(P, R_i) = p((x, y), R_i) \tag{16.182}$$

the probability that a random walk starting at $P(x, y)$ will terminate at the boundary point R_i, then one gets

$$p(R_i, R_i) = 1, \quad p(R_i, R_j) = 0 \quad \text{for} \quad i \neq j \quad \text{and} \tag{16.183}$$

$$p((x, y), R_i) = \frac{1}{4}[p((x-1, y), R_i)+p((x+1, y), R_i)+p((x, y-1), R_i)+p((x, y+1), R_i)]. \tag{16.184}$$

The equation (16.184) is a difference equation for $p((x, y), R_i)$. If starting n random walks from the point $P(x, y)$, from which m_i terminates at R_i $(m_i \leq n)$, then

$$p((x, y), R_i)) \approx \frac{m_i}{n}. \tag{16.185}$$

The equation (16.185) gives an approximate solution of the differential equation (16.180a) with the boundary condition (16.181). The boundary condition (16.180b) will be fulfilled if substituting

$$v(P) = v(x, y) = \sum_{i=1}^{N} f(R_i)p((x, y), R_i), \tag{16.186}$$

because of (16.184), $v(R_j) = \sum\limits_{i=1}^{N} f(R_i)p(R_j, R_i) = f(R_j)$.

To calculate $v(x, y)$ (16.184) is multiplied by $f(R_i)$. Summation yields the following difference equation for $v(x, y)$:

$$v(x, y) = \frac{1}{4}[v(x-1, y) + v(x+1, y) + v(x, y-1) + v(x, y+1)]. \tag{16.187}$$

Starting n random walks from an interior point $P(x, y)$, and among them m_j terminate at the boundary point R_i $(i = 1, 2, \ldots, N)$, then an approximate value is obtained at the point $P(x, y)$ of the boundary value problem (16.180a,b) by

$$v(x, y) \approx \frac{1}{n}\sum_{i=1}^{n} m_i f(R_i). \tag{16.188}$$

16.3.5.5 Further Applications of the Monte Carlo Method

Monte Carlo methods as stochastic simulation, sometimes called *methods of statistical experiments*, are used in many different areas. For examples:

- Nuclear techniques: Neutrons passing through material layers.
- Communication: Separating signals and noise.
- Operations research: Queueing systems, process design, inventory control, service.

For further details of these problem areas see for example [16.15].

16.4 Calculus of Errors

Every scientific measurement, giving certain numerical quantities – regardless of the care with which the measurements are made – is always subject to errors and uncertainties. There are observational errors, errors of the measuring method, instrumental errors and often errors arising from the inherent random nature of the phenomena being measured. Together they compose the *measurement error*.

All measurement errors arising during the measuring process are called *deviations*. As a consequence a measured quantity represented by a number of significant digits can be given only with a rounding error, i.e., with a certain statistical error, which is called the *uncertainty* of the result.

1. The deviations of the measuring process should be kept as small as possible. On this basis it is to be evaluated the possible best approximation, what can be done with the help of *smoothing methods* which have their origin in the Gaussian least squares method.

2. The uncertainties have to be estimated as well as possible, what can be done with the help of *methods of mathematical statistics*.

Because of the random character of the measuring results they can be considered as statistical samples (see 16.2.3, **1.**, p. 814) with its probability distribution, whose parameters contain the desired information. In this sense, measurement errors can be seen as sampling errors.

16.4.1 Measurement Error and its Distribution

16.4.1.1 Qualitative Characterization of Measurement Errors

Qualifying the measurement errors by their causes, the following three types of errors are distinguished:

1. Rough errors are caused by inaccurate readings or confusion; they are excludable.

2. Systematic measurement errors are caused by inaccurately scaled measuring devices and by the method of measuring, where the method of reading the data and also the measured error of the measurement system can play a role. They are not always avoidable.

3. Statistical or random measurement errors can arise from random changes of the measuring conditions that are difficult or impossible to control and also by certain random properties of the events observed.

In the theory of measurement errors usually it is assumed that the rough errors and the systematic measurement errors are excludable, so that only the statistical properties and the random measurement errors are included into the calculation of the errors.

16.4.1.2 Density Function of the Measurement Error

1. Measurement Protocol

To calculate the characterization of the uncertainties it must be supposed that the measured results are listed in a *measurement record* as a *prime notation* and that the relative frequencies or the density function $f(x)$ or the cumulative frequencies or the distribution function $F(x)$ (see 16.3.2.1, p. 832) of the uncertain values are available. The realization of the random variable X under consideration is denoted by x.

2. Error Density Function

Special assumptions about the properties of the measurement error result in certain special properties of the density function of the error distribution:

1. Continuous Density Function Since the random measurement errors can take any value in a certain interval, they are described by a continuous density function $f(x)$.

2. Even Density Function If measurement errors with the same absolute value but with different signs are equally likely, then the density function is an even function: $f(-x) = f(x)$.

3. Monotonically Decreasing Density Function If a measuring error with larger absolute value is less likely than an error with smaller absolute value, then the density function $f(x)$ is monotonically decreasing for $x > 0$.

4. Finite Expected Value The expected value of the absolute value of the error must be finite:

$$E(|X|) = \int_{-\infty}^{\infty} |x| f(x)\, dx < \infty. \tag{16.189}$$

Different properties of the errors result in different types of density functions.

3. Normal Distribution of the Error

1. Density Function and Distribution Function In most practical cases it can be supposed that the distribution of the measurement error is a normal distribution with expected value $\mu = 0$ and variance σ^2, i.e., the density function $f(x)$ and the distribution function $F(x)$ of the measurement error are:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} \quad (16.190a) \qquad \text{and } F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2\sigma^2}}\, dt = \Phi\left(\frac{x}{\sigma}\right). \quad (16.190b)$$

Here $\Phi(x)$ is the distribution function of the standard normal distribution (see (16.73a), p. 819, and **Table 21.17**, p. 1133). In the case of (16.190a,b) one speaks about *normal errors*.

2. Geometrical Representation The density function (16.190a) is represented in **Fig. 16.18a** with inflection points and points at the center of gravity, and its behavior is shown in **Fig. 16.18b** when the variance changes. The inflection points are at the abscissa values $\pm\sigma$; the centers of gravity of the half-areas are at $\pm\eta$. The maximum of the function is at $x = 0$ and it is $1/(\sigma\sqrt{2\pi})$. The curve widens as σ^2 increases, the area under the curve always equals one. This distribution shows that small errors occur often, large errors only seldom.

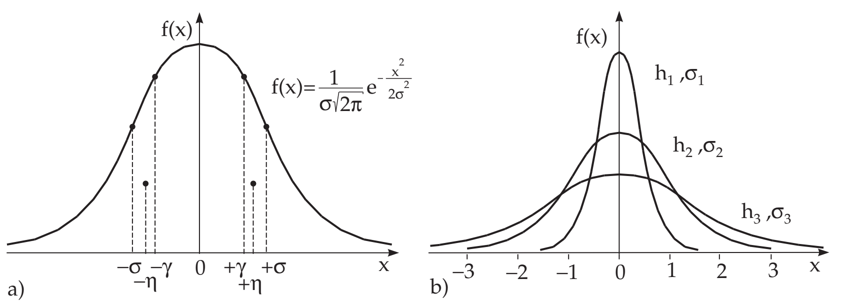


Figure 16.18

4. Parameters to Characterize the Normally Distributed Error

Beside the variance σ^2 or the standard deviation σ which is also called the *mean square error* or *standard error*, there are other parameters to characterize the normally distributed error, such as the measure of accuracy h, the *average error* or *mean error* η, and the *probable error* γ.

1. Measure of Accuracy Beside the variance σ^2, the *measure of accuracy*

$$h = \frac{1}{\sigma\sqrt{2}} \tag{16.191}$$

is used to characterize the width of the normal distribution. A narrower Gauss curve results in better accuracy **(Fig. 16.18b)**. Replacing σ by the experimental value $\tilde{\sigma}$ or $\tilde{\sigma}_x$ obtained from the measured values, the measure of accuracy characterizes the accuracy of the measurement method.

2. Average or Mean Error The expected value η of the absolute value of the error is defined as

$$\eta = E(|X|) = 2\int_0^\infty x f(x)\,dx = \sqrt{\frac{2}{\pi}}\sigma\,. \tag{16.192}$$

3. Probable Error The bound γ of the absolute value of the error with the property

$$P(|X| \le \gamma) = \frac{1}{2} \tag{16.193a}$$

is called the *probable error*. It implies that

$$\int_{-\gamma}^{+\gamma} f(x)\,dx = 2\Phi\left(\frac{\gamma}{\sigma}\right) - 1 = \sqrt{\frac{2}{\pi}}\int_0^{\gamma/\sigma} e^{-t^2/2}\,dt = \frac{1}{2}\,, \tag{16.193b}$$

where $\Phi(x)$ is the distribution function of the standard normal distribution. The condition (16.193b) is a non-linear equation which can be solved approximately by the help of a computer algebra system to get γ/σ. The result is

$$\frac{\gamma}{\sigma} \approx 0.6745\,. \tag{16.193c}$$

4. Given Error Bounds If an upper bound $a > 0$ of an error is given, then with (16.193b) can be calculated the probability that the error is in the interval $[-a, a]$:

$$P(|X| \le a) = 2\Phi\left(\frac{a}{\sigma}\right) - 1. \tag{16.194}$$

5. Relations between Standard Deviation, Average Error, Probable Error, and Accuracy If the error has a normal distribution, then the following relations hold using (16.193c):

$$\eta = \sqrt{\frac{2}{\pi}}\sigma\,, \quad (16.195a) \qquad \gamma = 0.6745\,\sigma\,, \quad (16.195b) \qquad h = \frac{1}{2\sqrt{\sigma}}\,. \quad (16.195c)$$

16.4.1.3 Quantitative Characterization of the Measurement Error

1. True Value and its Approximations

The true value x_w of a measurable quantity is usually unknown. Therefore the expected value of the random variables, whose realizations are the measured values x_i $(i = 1, 2, \ldots, n)$, is chosen as an estimated value of x_w. Consequently, the following means can be considered as an approximation of x_w.

1. Arithmetical Mean

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad (16.196a) \qquad \text{or} \qquad \overline{x} = \sum_{j=1}^{k} h_j \overline{x}_j\,, \quad (16.196b)$$

if the measured values are distributed into k classes with absolute frequencies h_j and class means $\overline{x}_j$ $(j = 1, 2, \ldots, k)$.

2. Weighted Mean

$$\overline{x}^{(g)} = \sum_{i=1}^{n} g_i x_i \Big/ \sum_{i=1}^{n} g_i. \tag{16.197}$$

Here the single measured values are weighted by the *weighting factors* g_i $(g_i > 0)$ (see 16.4.1.6, **1.**, p. 854).

2. Error of a Single Measurement in a Measurement Sequence

1. The True Error of a Single Measurement in a Measurement Sequence is the difference between the true value x_w and the measuring result. Because this is usually unknown, the *true error*

ε_i of the i-th measurement with the result x_i is also unknown:

$$\varepsilon_i = x_w - x_i. \tag{16.198}$$

2. The Mean Error or Apparent Error of a Single Measurement in a Measurement Sequence is the difference of the arithmetical mean and the measurement result x_i:

$$v_i = \overline{x} - x_i. \tag{16.199}$$

3. Mean Square Error of a Single Measurement or Standard Error of a Single Measurement Since the expected value of the sum of the true errors ε_i and the expected value of the sum of the apparent errors v_i of n measurements are zero (independently of how large they are), they are characterized by the sum of the error squares:

$$\varepsilon^2 = \sum_{i=1}^{n} {\varepsilon_i}^2, \tag{16.200a}$$

$$v^2 = \sum_{i=1}^{n} {v_i}^2. \tag{16.200b}$$

From a practical point of view only the value of (16.200b) is interesting, since only the values of v_i can be determined from the measuring process. Therefore, the mean square error of a single measurement of a measurement sequence is defined by

$$\tilde{\sigma} = \sqrt{\sum_{i=1}^{n} {v_i}^2 \Big/ (n-1)}. \tag{16.200c}$$

The value $\tilde{\sigma}$ is an approximation of the standard deviation σ of the error distribution.
One gets for $\tilde{\sigma} = \sigma$ in the case of normally distributed error:

$$P(|\varepsilon| \le \tilde{\sigma}) = 2\Phi(1) - 1 = 0.6826. \tag{16.200d}$$

That is: The probability that the absolute value of the true value does not exceed σ, is about 68%.

4. The Probable Error of a single measurement is the number γ, for which

$$P(|\varepsilon| \le \gamma) = \frac{1}{2}. \tag{16.201a}$$

That is: The probability that the absolute value of the error does not exceed γ, is 50%. The abscissae $\pm\gamma$ divide the area of the left and the right parts under the density function into two equal parts (**Fig. 16.18a**).
In the case of a normally distributed error in accordance to (16.193c) holds

$$\gamma \approx 0.6745\,\sigma \approx 0.6745\,\tilde{\sigma} = \tilde{\gamma}. \tag{16.201b}$$

5. Average Error of a single measurement is the number η, which is the expected value of the absolute value of the error:

$$\eta = E(|\varepsilon|) = \int_{-\infty}^{\infty} |x| f(x)\,dx. \tag{16.202a}$$

In the case of a normally distributed error follows $\eta = \sqrt{\frac{2}{\pi}}\sigma$ and

$$P(|\varepsilon| \le \eta) = 2\Phi\left(\frac{\eta}{\sigma}\right) - 1 = 2\Phi\left(\sqrt{\frac{2}{\pi}}\right) - 1 \approx 0.5751 : \tag{16.202b}$$

The probability that the error does not exceed the value η is about 57.6 %. The centers of gravity of the left and right areas under the density function (**Fig. 16.18a**) are at abscissae $\pm\eta$. In the case of normally distributed errors:

$$\eta = \sqrt{\frac{2}{\pi}}\sigma \approx \sqrt{\frac{2}{\pi}}\tilde{\sigma} = \tilde{\eta}. \tag{16.202c}$$

3. Error of the Arithmetical Mean of a Measurement Sequence

The error of the arithmetical mean $\overline{x}$ of a measurement sequence is given by the errors of the single measurement:

1. Mean Square Error or Standard Deviation

$$\tilde{\sigma}_{AM} = \sqrt{\sum_{i=1}^{n} {v_i}^2 \Big/ [n(n-1)]} = \frac{\tilde{\sigma}}{\sqrt{n}}\,. \tag{16.203}$$

2. Probable Error

$$\tilde{\gamma}_{AM} \approx 0.6745 \sqrt{\sum_{i=1}^{n} {v_i}^2 \Big/ [n(n-1)]} = 0.6745 \frac{\tilde{\sigma}}{\sqrt{n}}\,. \tag{16.204}$$

3. Average Error

$$\tilde{\eta}_{AM} \approx 0.7979 \sqrt{\sum_{i=1}^{n} {v_i}^2 \Big/ [n(n-1)]} = 0.7979 \frac{\tilde{\sigma}}{\sqrt{n}}\,. \tag{16.205}$$

4. Accessible Level of Error Since the three types of errors defined above (16.203)–(16.205) are directly proportional to the corresponding error of the single measurement (16.200c), (16.201b) and (16.202c) and they are proportional to the reciprocal of the square root of n, it is not reasonable to increase the number of the measurements after a certain value. It is more efficient to improve the accuracy h of the measuring method (16.191).

4. Absolute and Relative Errors

1. Absolute Uncertainty, Absolute Error The uncertainty of the results of measurement is characterized by errors ε_i, v_i, σ_i, γ_i, η_i, or ε, v, σ, γ, η, which measure the reliability of the measurements. The notion of the *absolute uncertainty*, given as the *absolute error*, is meaningful for all these types of errors and for the calculation of error propagation (see 16.4.2, p. 854). They have the same dimension as the measured quantity.
The word "absolute" error is introduced to avoid confusion with the notion of relative error. Often the notation Δx_i or Δx is used. The word "absolute" has a different meaning from the notion of absolute value: It refers to the numerical value of the measured quantity (e.g., length, weight, energy), without restriction of its sign.

2. Relative Uncertainty, Relative Error The *relative uncertainty*, given by the *relative error*, is a measure of the quality of the method of measurement with respect to the numerical value of the measured quantity. In contrast to the absolute error, the *relative error* has no dimension, because it is the quotient of the absolute error and the numerical value of the measured quantity. If this value is not known, one replaces it by the mean value of the quantity x:

$$\delta x_i = \frac{\Delta x_i}{x} \approx \frac{\Delta x_i}{\overline{x}}\,. \tag{16.206a}$$

The relative error is given mostly as a percentage and it is also called the *percentage error*:

$$\delta x_i/\% = \delta x_i \cdot 100\,\%. \tag{16.206b}$$

5. Absolute and Relative Maximum Error

1. Absolute Maximum Error If the quantity z is to be determined and is z a function of the measured quantities $x_1, x_2, \ldots, x_n$, i.e., $z = f(x_1, x_2, \ldots, x_n)$, then the resulting error must be calculated taking also the function f into consideration. There are two different ways to examine errors. The first approach is that statistical error analysis is applied by smoothing the data values using the least squares method ($\min \sum (z_i - z)^2$). In the second approach, an upper bound $\Delta z_{\max}$ is determined for the absolute error of the quantities. With n independent variables x_i, holds:

$$\Delta z_{\max} = \sum_{i=1}^{n} \left| \frac{\partial}{\partial x_i} f(x_1, x_2, \ldots, x_n) \right| \Delta x_i, \tag{16.207}$$

where the mean value $\overline{x}_i$ should be substituted for x_i.

2. Relative Maximum Error The relative maximum error is the absolute maximum error divided by the numerical value of the measured value (mostly by the mean of z):

$$\delta z_{\max} = \frac{\Delta z_{\max}}{z} \approx \frac{\Delta z_{\max}}{\overline{z}}\,. \tag{16.208}$$

16.4.1.4 Determining the Result of a Measurement with Bounds on the Error

A realistic interpretation of a measurement result is possible only if the expected error is also given; error estimations and bounds are components of measurement results. It must be clear from the data what is the type of the error, what is the confidence interval and what is the significance level.

1. Defining the Error The result of a single measurement is required to be given in the form

$$x = x_i \pm \Delta x \approx x_i \pm \tilde{\sigma}, \tag{16.209a}$$

and the result of the mean has the form

$$x = \overline{x} \pm \Delta x_{AM} \approx \overline{x} \pm \tilde{\sigma}_{AM}. \tag{16.209b}$$

Here for Δx the most often used standard deviation $\tilde{\sigma}$ is applied. $\tilde{\gamma}$ and $\tilde{\eta}$ could also be used.

2. Prescription of Arbitrary Confidence Limits The quantity $T = \dfrac{X - x_w}{\tilde{\sigma}_{AM}}$ has a t distribution (16.97b) with $f = n - 1$ degrees of freedom in the case of a population with a distribution $N(\mu, \sigma^2)$ according to (16.96). For a required significance level α or for an *acceptance probability* $S = 1 - \alpha$ the confidence limits for the unknown quantity $x_w = \mu$ with the t quantile $t_{\alpha/2;f}$ are

$$\mu = \overline{x} \pm t_{\alpha/2;f} \cdot \tilde{\sigma}_{AM}. \tag{16.210}$$

That is, the true value x_w is in the interval given by these limits with a probability $S = 1 - \alpha$.
Mostly it is of interest keeping the size n of the measurement sequence at its lowest possible level. The length $2t_{\alpha/2;f}\tilde{\sigma}_{AM}$ of the confidence interval decreases for a smaller value of $1 - \alpha$ and also for a larger number n of measurements. Since $\tilde{\sigma}_{AM}$ decreases proportionally to $1/\sqrt{n}$ and the quantile $t_{\alpha/2;f}$ with $f = n - 1$ degrees of freedom also decreases proportionally to $1/\sqrt{n}$ (for values of n between 5 and 10, see **Table 21.20**, p. 1138) the length of the confidence interval decreases proportionally to $1/n$ for such values of n.

16.4.1.5 Error Estimation for Direct Measurements with the Same Accuracy

If there is the same variance σ_i for all n measurements, one talks about measurements with the same accuracy h = const. In this case, the least squares method results in the error quantities given in (16.200c), (16.201b), and (16.202b).

■ Determine the final result for the measurement sequence given in the following table which contains $n = 10$ direct measurements with the same accuracy.

x_i	1.592	1.581	1.574	1.566	1.603	1.580	1.591	1.583	1.571	1.559
$v_i \cdot 10^3$	− 12	− 1	+ 6	+ 14	− 23	0	− 11	− 3	+ 9	+ 21
${v_i}^2 \cdot 10^6$	144	1	36	196	529	0	121	9	81	441

$\overline{x} = 1.580$, $\tilde{\sigma} = \sqrt{\sum\limits_{i=1}^{n} {v_i}^2/(n-1)} = 0.0131$, $\tilde{\sigma}_{AM} = \tilde{\sigma}/\sqrt{n} = 0.004$.

Final result: $x = \overline{x} \pm \tilde{\sigma}_{AM} = 1.580 \pm 0.004$.

16.4.1.6 Error Estimation for Direct Measurements with Different Accuracy

1. Weighted Measurements

If the direct measurement results x_i are obtained from different measuring methods or they represent means of single measurements, which belong to the same mean $\overline{x}$ with different variances $\tilde{\sigma_i}^2$, then a *weighted mean*

$$\overline{x}^{(g)} = \sum_{i=1}^{n} g_i x_i \Big/ \sum_{i=1}^{n} g_i \tag{16.211}$$

is calculated, where g_i is defined as

$$g_i = \frac{\tilde{\sigma}^2}{\tilde{\sigma_i}^2}\,. \tag{16.212}$$

Here $\tilde{\sigma}$ is an arbitrary positive value, mostly the smallest $\tilde{\sigma_i}$. It serves as a weight unit of the deviations, i.e., for $\tilde{\sigma_i} = \tilde{\sigma}$ it is $g_i = 1$. It follows from (16.210) that a larger weight of a measurement results in a smaller deviation $\tilde{\sigma_i}$.

2. Standard Deviations

The standard deviation of the weight unit is estimated as

$$\tilde{\sigma}^{(g)} = \sqrt{\sum_{i=1}^{n} g_i {v_i}^2 \Big/ (n-1)}\,. \tag{16.213}$$

It must be sure that $\tilde{\sigma}^{(g)} < \tilde{\sigma}$. In the opposite case, if $\tilde{\sigma}^{(g)} > \tilde{\sigma}$, then there are x_i values which have systematic deviations.

The standard deviation of the single measurement is

$$\tilde{\sigma_i}^{(g)} = \frac{\tilde{\sigma}^{(g)}}{\sqrt{g_i}} = \frac{\tilde{\sigma}^{(g)}}{\tilde{\sigma}}\tilde{\sigma_i}\,, \tag{16.214}$$

where $\tilde{\sigma_i}^{(g)} < \tilde{\sigma_i}$ can be expected.

The standard deviation of the weighted mean is:

$$\tilde{\sigma}_{AM}^{(g)} = \tilde{\sigma}^{(g)} \Big/ \sqrt{\sum_{i=1}^{n} g_i} = \sqrt{\sum_{i=1}^{n} g_i {v_i}^2 \Big/ \left((n-1)\sum_{i=1}^{n} g_i\right)}. \tag{16.215}$$

3. Error Description

The error can be described as it is represented in 16.4.1.4, p. 853, either by the definition of the error or by the t quantile with f degrees of freedom.

■ The final results of measurement sequences ($n = 5$) with different means $\overline{x}_i$ $(i = 1, 2, \ldots, 5)$ and with different standard deviations $\tilde{\sigma}_{AM_i}$ are given in **Table 16.6**.
Calculating $(\overline{x}_i)_m = 1.5830$ and choosing $x_0 = 1.585$ and $\tilde{\sigma} = 0.009$ with $z_i = \overline{x}_i - x_0$, $g_i = \tilde{\sigma}^2/\tilde{\sigma_i}^2$ one gets $\overline{z} = -0.0036$ and $\overline{x} = x_0 + \overline{z} = 1.582$. The standard deviation is $\tilde{\sigma}^{(g)} = \sqrt{\sum_{i=1}^{n} g_i {v_i}^2 \Big/ (n-1)} = 0.0088 < \tilde{\sigma}$ and $\tilde{\sigma_x} = \tilde{\sigma}_{AM} = 0.0027$. The final result is $x = \overline{x} \pm \tilde{\sigma_x} = 1.585 \pm 0.0027$.

16.4.2 Error Propagation and Error Analysis

Measured quantities appear in final results often in the form of functional dependencies. One talks about *error propagation*. If the error is small, then a Taylor expansion with respect to the error can be

Table 16.6 Error description of a measurement sequence

$\overline{x}_i$	$\tilde{\sigma}_{AM_i}$	$\tilde{\sigma}^2_{AM_i}$	g_i	z_i	g_iz_i	${z_i}^2$	$g_i{z_i}^2$
1.573	0.010	$1.0{\cdot}10^{-4}$	0.81	$-1.2{\cdot}10^{-2}$	$-9.7{\cdot}10^{-3}$	$1.44{\cdot}10^{-4}$	$1.16{\cdot}10^{-4}$
1.580	0.004	$1.6{\cdot}10^{-5}$	5.06	$-5.0{\cdot}10^{-3}$	$-2.5{\cdot}10^{-2}$	$2.50{\cdot}10^{-5}$	$1.26{\cdot}10^{-4}$
1.582	0.005	$2.5{\cdot}10^{-5}$	3.24	$-3.0{\cdot}10^{-3}$	$-9.7{\cdot}10^{-3}$	$9.0{\cdot}10^{-6}$	$2.91{\cdot}10^{-5}$
1.589	0.009	$8.1{\cdot}10^{-5}$	1.00	$+4.0{\cdot}10^{-3}$	$4.0{\cdot}10^{-3}$	$1.6{\cdot}10^{-5}$	$1.6{\cdot}10^{-5}$
1.591	0.011	$1.21{\cdot}10^{-4}$	0.66	$+6.0{\cdot}10^{-3}$	$3.9{\cdot}10^{-3}$	$3.6{\cdot}10^{-5}$	$2.37{\cdot}10^{-5}$
$(\overline{x}_i)_m$	$\tilde{\sigma}$		$\sum\limits_{i=1}^{n} g_i$		$\sum\limits_{i=1}^{n} g_iz_i$		$\sum\limits_{i=1}^{n} g_i{z_i}^2$
$=1.583$	$=0.009$		$=10.7$		$=3.6{\cdot}10^{-2}$		$=3.1{\cdot}10^{-4}$

used neglecting the terms of second and higher order.

16.4.2.1 Gauss Error Propagation Law

1. Problem Formulation

Suppose it is to determine the numerical value and the error of a quantity z given by the function $z = f(x_1, x_2, \ldots, x_k)$ of the independent variables x_j $(j = 1, 2, \ldots, k)$. The mean value $\overline{x}_j$ obtained from n_j measured values is considered as realizations of the random variable x_j, with variance ${\sigma_j}^2$. The question is how the errors of the variables affect the function value $f(x_1, x_2, \ldots, x_k)$. It is assumed that the function $f(x_1, x_2, \ldots, x_k)$ is differentiable and its variables are stochastically independent. However they may follow any type of distribution with different variances ${\sigma_j}^2$.

2. Taylor Expansion

Since the error represents relatively small changes of the independent variables, the function $f(x_1, x_2, \ldots, x_k)$ can be approximated in the neighborhood of the mean $\overline{x}_j$ by the linear part of its Taylor expansion with the coefficients a_j, so its error Δf is:

$$\Delta f = f(x_1, x_2, \ldots, x_k) - f(\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_k), \tag{16.216a}$$

$$\Delta f \approx df = \frac{\partial f}{\partial x_1}dx_1 + \frac{\partial f}{\partial x_2}dx_2 + \cdots + \frac{\partial f}{\partial x_k}dx_k = \sum_{j=1}^{k} \frac{\partial f}{\partial x_j}dx_j = \sum_{j=1}^{k} a_j dx_j, \tag{16.216b}$$

where the partial derivatives $\partial f/\partial x_j$ are taken at $(\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_k)$.
The variance of the function is

$${\sigma_f}^2 = {a_1}^2{\sigma_{x_1}}^2 + {a_2}^2{\sigma_{x_2}}^2 + \cdots + {a_k}^2{\sigma_{x_k}}^2 = \sum_{j=1}^{k} {a_j}^2{\sigma_{x_j}}^2. \tag{16.217}$$

3. Approximation of the Variance σ_f^2

Since the variances of the independent variables x_j are unknown, they can be approximated by the variance of their mean, which is determined from the measured values x_{jl} $(l = 1, 2, \ldots, n_l)$ of the single variables as follows:

$$\tilde{\sigma}^2_{\overline{x}_j} = \frac{\sum\limits_{l=1}^{n_j} (x_{jl} - \overline{x}_j)^2}{n_j(n_j - 1)}. \tag{16.218}$$

Using these values one can form an approximation of ${\sigma_f}^2$:

$$\tilde{\sigma}^2_f = \sum_{j=1}^{k} {a_j}^2 \tilde{\sigma}^2_{\overline{x}_j}. \tag{16.219}$$

The formula (16.219) is called the *Gauss error propagation law.*

4. Special Cases

1. Linear Case An often occurring case is the summation of the values of the errors of linearly occurring error quantities with $a_j = 1$:

$$\tilde{\sigma}_f = \sqrt{\tilde{\sigma}_1^2 + \tilde{\sigma}_2^2 + \cdots + \tilde{\sigma}_k^2}. \tag{16.220}$$

■ The pulse length is to be measured at the output of a pulse amplifier of a detector channel for spectrometry of radiation, whose error can be deduced for three components:
1. statistical energy distribution of the radiation of the part passing through the spectometer with an energy E_0, which is characterized by $\tilde{\sigma}_{\text{Str}}$,
2. statistical interference processes in the detector with $\tilde{\sigma}_{\text{Det}}$,
3. electronic noice of the amplifier of the detector impulse $\tilde{\sigma}_{\text{el}}$.
The total pulse length has the error

$$\tilde{\sigma}_f = \sqrt{\tilde{\sigma}_{\text{Str}}^2 + \tilde{\sigma}_{\text{Det}}^2 + \tilde{\sigma}_{\text{el}}^2}. \tag{16.221}$$

2. Power Rule The variables x_j often occur in the following form:

$$z = f(x_1, x_2, \ldots x_k) = a{x_1}^{b_1} \cdot {x_2}^{b_2} \ \ldots\ {x_k}^{b_k}. \tag{16.222}$$

By logarithmic differentiation the relative error is

$$\frac{df}{f} = b_1 \frac{dx_1}{x_1} + b_2 \frac{dx_2}{x_2} + \cdots + b_k \frac{dx_k}{x_k}, \tag{16.223}$$

from which by the error propagation law follows for the mean relative error:

$$\frac{\tilde{\sigma}_f}{f} = \sqrt{\sum_{j=1}^{k} \left(b_j \frac{\tilde{\sigma}_{\overline{x}_j}}{\overline{x}_j} \right)^2}. \tag{16.224}$$

■ Suppose the function $f(x_1, x_2, x_3)$ has the form $f(x_1, x_2, x_3) = \sqrt{x_1}{x_2}^2{x_3}^3$, and the standard deviations are σ_{x_1}, σ_{x_2} and σ_{x_3}. The relative error is then

$$\delta z = \frac{\tilde{\sigma}_f}{f} = \sqrt{\left(\frac{1}{2}\frac{\tilde{\sigma}_{\overline{x}_1}}{\overline{x}_1}\right)^2 + \left(2\frac{\tilde{\sigma}_{\overline{x}_2}}{\overline{x}_2}\right)^2 + \left(3\frac{\tilde{\sigma}_{\overline{x}_3}}{\overline{x}_3}\right)^2}.$$

5. Difference to the Maximum Error

Declaring the absolute or relative maximal error (16.207), (16.208) means that no *smoothing* for the values of the measurement is been made. For the determination of the relative or absolute error with the error propagation laws (16.219) or (16.222), smoothing between the measurement values x_j means that a confidence interval for a previously given level is to be determined. This procedure is given in 16.4.1.4, p. 853.

16.4.2.2 Error Analysis

The general analysis of error propagation in the calculations of a function $\varphi(x_i)$, when quantities of higher order are neglected, is called error analysis. In the framework of the theory of error analysis one investigates using an algorithm, how an input error Δx_i affects the value of $\varphi(x_i)$. In this context one also talks about differential error analysis.

In numerical mathematics, error analysis means the investigation of the affect of errors of methods, of rounding, and of input errors to the final result (see [19.24]).

17 Dynamical Systems and Chaos

17.1 Ordinary Differential Equations and Mappings

17.1.1 Dynamical Systems

17.1.1.1 Basic Notions

1. The Notion of Dynamical Systems and Orbits

A *dynamical system* is a mathematical object to describe the development of a physical, biological or another system from real life depending on time. It is defined by a *phase space* M, and by a one-parameter family of mappings $\varphi^t: M \to M$, where t is the parameter (the *time*). In the following discussion the phase space is often $\mathbb{R}^n$, a subset of it, or a metric space. The time parameter t is from $\mathbb{R}$ (*time continuous system*) or from $\mathbb{Z}$ or from $\mathbb{Z}_+$ (*time discrete system*). Furthermore, it is required for arbitrary $x \in M$ that

a) $\varphi^0(x) = x$ and

b) $\varphi^t(\varphi^s(x)) = \varphi^{t+s}(x)$ for all t, s. The mapping φ^1 is denoted briefly by φ.

Hereafter, the time set is denoted by Γ, hence, $\Gamma = \mathbb{R}, \Gamma = \mathbb{R}_+, \Gamma = \mathbb{Z}$ or $\Gamma = \mathbb{Z}_+$. If $\Gamma = \mathbb{R}$, then the dynamical system is also called a *flow*; if $\Gamma = \mathbb{Z}$ or $\Gamma = \mathbb{Z}_+$, then the dynamical system is *discrete*. In case $\Gamma = \mathbb{R}$ and $\Gamma = \mathbb{Z}$, the properties **a)** and **b)** are satisfied for every $t \in \Gamma$, so the inverse mapping $(\varphi^t)^{-1} = \varphi^{-t}$ also exists, and these systems are called *invertible* dynamical systems.

If the dynamical system is not invertible, then $\varphi^{-t}(A)$ means the pre-image of A with respect to φ^t, for an arbitrary set $A \subset M$ and arbitrary $t > 0$, i.e., $\varphi^{-t}(A) = \{x \in M: \varphi^t(x) \in A\}$. If the mapping $\varphi^t: M \to M$ is continuous or k times continuously differentiable for every $t \in \Gamma$ (here $M \subset \mathbb{R}^n$), then the dynamical system is called *continuous* or C^k*-smooth*, respectively.

For an arbitrary fixed $x \in M$, the mapping $t \longmapsto \varphi^t(x)$, $t \in \Gamma$, defines a *motion* of the dynamical system starting from x at time $t = 0$. The image $\gamma(x)$ of a motion starting at x is called the *orbit* (or the *trajectory*) through x, namely $\gamma(x) = \{\varphi^t(x)\}_{t\in\Gamma}$. Analogously, the *positive semiorbit* through x is defined by $\gamma^+(x) = \{\varphi^t(x)\}_{t\geq 0}$ and, if $\Gamma \neq \mathbb{R}_+$ or $\Gamma \neq \mathbb{Z}_+$, then the *negative semiorbit* through x is defined by $\gamma^-(x) = \{\varphi^t(x)\}_{t\leq 0}$.

The orbit $\gamma(x)$ is a *steady state* (also *equilibrium point* or *stationary point*) if $\gamma(x) = \{x\}$, and it is T*-periodic* if there exists a $T \in \Gamma$, $T > 0$, such that $\varphi^{t+T}(x) = \varphi^t(x)$ for all $t \in \Gamma$, and $T \in \Gamma$ is the smallest positive number with this property. The number T is called the *period*.

2. Flow of a Differential Equation

Consider the *ordinary linear planar differential equation*

$$\dot{x} = f(x), \tag{17.1}$$

where $f: M \to \mathbb{R}^n$ (*vector field*) is an r-times continuously differentiable mapping and $M = \mathbb{R}^n$ or M is an open subset of $\mathbb{R}^n$. From now on, the Euclidean norm $\|\cdot\|$ is used in $\mathbb{R}^n$, i.e., for arbitrary $x \in \mathbb{R}^n$, $x = (x_1, \ldots, x_n)$, its norm is $\|x\| = \sqrt{\sum_{i=1}^{n} x_i^2}$. If the mapping f is written componentwise $f = (f_1, \ldots, f_n)$, then (17.1) is a system of n scalar differential equations $\dot{x}_i = f_i(x_1, \ldots, x_n)$, $i = 1, 2, \ldots, n$.

The Picard–Lindelöf theorem on the local existence and uniqueness of solutions of differential equations and the theorem on the r*-times differentiability of solutions with respect to the initial values* (see [17.11]) guarantee that for every $x_0 \in M$, there exist a number $\varepsilon > 0$, a sphere $B_\delta(x_0) = \{x: \|x - x_0\| < \delta\}$ in M and a mapping $\varphi: (-\varepsilon, \varepsilon) \times B_\delta(x_0) \to M$ such that:

1. $\varphi(\cdot,\cdot)$ is $(r+1)$-times continuously differentiable with respect to its first argument (time) and r-times continuously differentiable with respect to its second argument (phase variable);

I.N. Bronshtein et al., *Handbook of Mathematics*,
DOI 10.1007/978-3-662-46221-8_17

2. For every fixed $x \in B_\delta(x_0)$, $\varphi(\cdot, x)$ is the *locally unique solution* of (17.1) in the time interval $(-\varepsilon, \varepsilon)$ which starts from x at time $t = 0$, i.e., $\frac{\partial \varphi}{\partial t}(t,x) = \dot{\varphi}(t,x) = f(\varphi(t,x))$ holds for every $t \in (-\varepsilon, \varepsilon)$, $\varphi(0,x) = x$, and every other solution with initial point x at time $t = 0$ coincides with $\varphi(t,x)$ for all small $|t|$.

Suppose that every local solution of (17.1) can be extended uniquely to the whole of $\mathbb{R}$. Then there exists a mapping $\varphi\colon \mathbb{R} \times M \to M$ with the following properties:

1. $\varphi(0,x) = x$ for all $x \in M$.

2. $\varphi(t+s,x) = \varphi(t, \varphi(s,x))$ for all $t,\ s \in \mathbb{R}$ and all $x \in M$.

3. $\varphi(\cdot,\cdot)$ is continuously differentiable $(r+1)$ times with respect to its first argument and r times with respect to the second one.

4. For every fixed $x \in M$, $\varphi(\cdot,x)$ is a solution of (17.1) on the whole of $\mathbb{R}$.

Then the C^r-smooth flow generated by (17.1) can be defined by $\varphi^t := \varphi(t,\cdot)$. The motions $\varphi(\cdot,x)\colon \mathbb{R} \to M$ of a flow of (17.1) are called *integral curves.*

■ The equation

$$\dot{x} = \sigma(y-x), \quad \dot{y} = rx - y - xz, \quad \dot{z} = xy - bz \tag{17.2}$$

is called a *Lorenz system of convective turbulence* (see also 17.2.4.3, p. 887). Here $\sigma > 0$, $r > 0$ and $b > 0$ are parameters. A C^∞ flow on $M = \mathbb{R}^3$ corresponds to the Lorenz system .

3. Discrete Dynamical System

Consider the difference equation

$$x_{t+1} = \varphi(x_t), \tag{17.3}$$

which can also be written as an assignment $x \longmapsto \varphi(x)$. Here $\varphi\colon M \to M$ is a continuous or r times continuously differentiable mapping, where in the second case $M \subset \mathbb{R}^n$. If φ is invertible, then (17.3) defines an invertible discrete dynamical system through the iteration of φ, namely,

$$\varphi^t = \underbrace{\varphi \circ \cdots \circ \varphi}_{t\,\text{times},}, \text{ for } \quad t > 0, \qquad \varphi^t = \underbrace{\varphi^{-1} \circ \cdots \circ \varphi^{-1}}_{-t\,\text{times},}, \quad \text{for} \quad t < 0,\ \varphi^0 = id. \tag{17.4}$$

If φ is not invertible, then the mappings φ^t are defined only for $t \geq 0$. For the realization of φ^t see (5.74), p. 333.

■ **A:** The difference equation

$$x_{t+1} = \alpha x_t\,(1 - x_t), \quad t = 0, 1, \ldots \tag{17.5}$$

with parameter $\alpha \in (0,4]$ is called a *logistic equation*. Here $M = [0,1]$, and $\varphi\colon [0,1] \to [0,1]$ is the function $\varphi(x) = \alpha x(1-x)$ for a fixed α. Obviously, φ is infinitely many times differentiable, but not invertible. Hence (17.5) defines a non-invertible dynamical system.

■ **B:** The difference equation

$$x_{t+1} = y_t + 1 - a x_t^2, \quad y_{t+1} = b x_t, \quad t = 0, \pm 1, \ldots, \tag{17.6}$$

with parameters $a > 0$ and $b \neq 0$ is called a *Hénon mapping*. The mapping $\varphi\colon \mathbb{R}^2 \to \mathbb{R}^2$ corresponding to (17.6) is defined by $\varphi(x,y) = (y+1-ax^2, bx)$, it is infinitely many times differentiable and invertible.

4. Volume Contracting and Volume Preserving Systems

The invertible dynamical system $\{\varphi^t\}_{t \in \Gamma}$ on $M \subset \mathbb{R}^n$ is called *volume depressing* or *dissipative* (*volume preserving* or *conservative*), if the relation $\text{vol}(\varphi^t(A)) < \text{vol}(A)$ $(\text{vol}(\varphi^t(A)) = \text{vol}(A))$ holds for every set $A \subset M$ with a positive n-dimensional volume $\text{vol}(A)$ and every $t > 0\,(t \in \Gamma)$.

■ **A:** Let φ in (17.3) be a C^r*-diffeomorphism* (i.e.: $\varphi\colon M \to M$ is invertible, $M \subset \mathbb{R}^n$ open, φ and φ^{-1} are C^r-smooth mappings) and let $D\varphi(x)$ be the Jacobi matrix of φ in $x \in M$. The discrete system (17.3) is dissipative if $|\det D\varphi(x)| < 1$ for all $x \in M$, and conservative if $|\det D\varphi(x)| \equiv 1$ in M.

■ **B:** For the system (17.6) $D\varphi(x,y) = \begin{pmatrix} -2ax & 1 \\ b & 0 \end{pmatrix}$ and so $|\det D\varphi(x,y)| \equiv b$. Hence, (17.6) is dissipative if $|b| < 1$, and conservative if $|b| = 1$.
The Hénon mapping can be decomposed into three mappings **(Fig. 17.1)**: First, the initial domain is stretched and bent by the mapping $x' = x, y' = y + 1 - ax^2$ in a area-preserving way, then it is contracted in the direction of the x'-axis by $x'' = bx', y'' = y'$ (at $|b| < 1$), and finally it is reflected with respect to the line $y'' = x''$ by $x''' = y''$, $y''' = x''$.

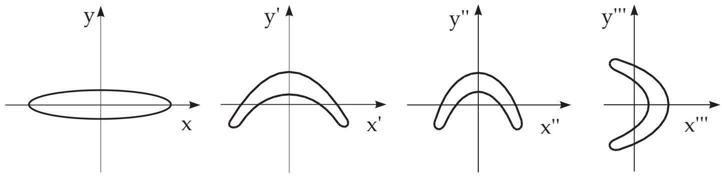


Figure 17.1

17.1.1.2 Invariant Sets

1. α- and ω-Limit Sets, Absorbing Sets

Let $\{\varphi^t\}_{t\in\Gamma}$ be a dynamical system on M. The set $A \subset M$ is *invariant under* $\{\varphi^t\}$, if $\varphi^t(A) = A$ holds for all $t \in \Gamma$, and *positively invariant under* $\{\varphi^t\}$, if $\varphi^t(A) \subset A$ holds for all $t \geq 0$ from Γ.
For every $x \in M$, the *ω-limit set* of the orbit passing through x is the set

$$\omega(x) = \{y \in M \colon \exists\, t_n \in \Gamma, \quad t_n \to +\infty, \quad \varphi^{t_n}(x) \to y \quad \text{as}\, n \to +\infty\}. \tag{17.7}$$

The elements of $\omega(x)$ are called *ω-limit points* of the orbit. If the dynamical system is invertible, then for every $x \in M$, the set

$$\alpha(x) = \{y \in M \colon \exists\, t_n \in \Gamma, \quad t_n \to -\infty, \quad \varphi^{t_n}(x) \to y \quad \text{as}\, n \to +\infty\} \tag{17.8}$$

is called the *α-limit set of the orbit passing through x;* the elements of $\alpha(x)$ are called the *α-limit points* of the orbit.
For many systems which are volume decreasing under the flow there exists a bounded set in phase space such that every orbit reaching it stays there as time increases. A bounded, open and connected set $U \subset M$ is called *absorbing* with respect to $\{\varphi^t\}_{t\in\Gamma}$, if $\varphi^t(\overline{U}) \subset U$ holds for all positive t from Γ. ($\overline{U}$ is the closure of U.)

■ Consider the system of differential equations

$$\dot{x} = -y + x\,(1 - x^2 - y^2), \quad \dot{y} = x + y\,(1 - x^2 - y^2) \tag{17.9a}$$

in the plane. Using the polar coordinates $x = r\cos\vartheta$, $y = r\sin\vartheta$, the solution of (17.9a) with initial state (r_0, ϑ_0) at time $t = 0$ has the form

$$r(t, r_0) = [1 + (r_0^{-2} - 1)\,e^{-2t}]^{-1/2}, \quad \vartheta(t, \vartheta_0) = t + \vartheta_0. \tag{17.9b}$$

This representation of the solution shows that the flow of (17.9a) has a periodic orbit with period 2π, which can be given in the form $\gamma((1,0)) = \{(\cos t, \sin t), t \in [0, 2\pi]\}$. The limit sets of an orbit through p are:

$$\alpha(p) = \begin{cases} (0,0), & \|\, p \,\| < 1, \\ \gamma((1,0)), & \|\, p \,\| = 1, \\ \emptyset, & \|\, p \,\| > 1 \end{cases} \quad \text{and} \quad \omega(p) = \begin{cases} \gamma((1,0)), & p \neq (0,0), \\ (0,0), & p = (0,0). \end{cases}$$

Every open sphere $B_r = \{(x,y) \colon x^2 + y^2 < r^2\}$ with $r > 1$ is an absorbing set for (17.9a).

2. Stability of Invariant Sets

Let A be an invariant set of the dynamical system $\{\varphi^t\}_{t\in\Gamma}$ defined on (M, ρ). The set A is called *stable*, if every neighborhood U of A contains another neighborhood $U_1 \subset U$ of A such that $\varphi^t(U_1) \subset U$ holds

for all $t > 0$. The set A, which is invariant under $\{\varphi^t\}$, is called *asymptotically stable* if it is stable and the following relations are satisfied:

$$\exists\, \Delta > 0 \quad \left.\begin{matrix} \forall x \in M \\ \operatorname{dist}(x, A) < \Delta \end{matrix}\right\}: \quad \operatorname{dist}(\varphi^t(x), A) \longrightarrow 0 \quad \text{for} \quad t \to +\infty. \tag{17.10}$$

Here, $\operatorname{dist}(x, A) = \inf\limits_{y \in A} \rho(x, y)$.

3. Compact Sets

Let (M, ρ) be a metric space. A system $\{U_i\}_{i \in \mathrm{I}}$ of open sets is called an *open covering* of M if every point of M belongs to at least one U_i. The metric space (M, ρ) is called *compact* if it is possible to choose finitely many $U_{i_1}, \ldots, U_{i_r}$ from every open covering $\{U_i\}_{i \in \mathrm{I}}$ of M such that $M = U_{i_1} \cup \cdots \cup U_{i_r}$ holds. The set $K \subset M$ is called compact if it is compact as a subspace.

4. Attractor, Domain of Attraction

Let $\{\varphi^t\}_{t \in \Gamma}$ be a dynamical system on (M, ρ) and A an invariant set for $\{\varphi^t\}$. Then $W(A) = \{x \in M\colon \omega(x) \subset A\}$ is called the *domain of attraction* of A. A compact set $\Lambda \subset M$ is called an *attractor* of $\{\varphi^t\}_{t \in \Gamma}$ on M if Λ is invariant under $\{\varphi^t\}$ and there is an open neighborhood U of Λ such that $\omega(x) = \Lambda$ for almost every (in the sense of Lebesgue measure) $x \in U$.

■ $\Lambda = \gamma((1, 0))$ is an attractor of the flow of (17.9a). Here $W(\Lambda) = \mathbb{R}^2 \setminus \{(0, 0)\}$. For some dynamical systems, a more general notion of an attractor makes sense. So, there are invariant sets Λ which have periodic orbits in every neighborhood of Λ which are not attracted by Λ, e.g., the Feigenbaum attractor. The set Λ may not be generated by a single limit set ω. A compact set Λ is called an *attractor in the sense of Milnor* of the dynamical system $\{\varphi^t\}_{t \in \Gamma}$ on M if Λ is invariant under $\{\varphi^t\}$ and the domain of attraction of Λ contains a set with positive Lebesgue measure.

17.1.2 Qualitative Theory of Ordinary Differential Equations

17.1.2.1 Existence of Flows, Phase Space Structure

1. Extensibility of Solutions

Besides the differential equation (17.1), which is called *autonomous*, there are differential equations whose right-hand side depends explicitly on the time and they are called *non-autonomous*:

$$\dot{x} = f(t, x). \tag{17.11}$$

Let $f\colon \mathbb{R} \times M \to M$ be a C^r-mapping with $M \subset \mathbb{R}^n$. By the new variable $x_{n+1} := t$, (17.11) can be interpreted as the autonomous differential equation $\dot{x} = f(x_{n+1}, x)$, $\dot{x}_{n+1} = 1$. The solution of (17.11) starting from x_0 at time t_0 is denoted by $\varphi(\cdot, t_0, x_0)$. In order to show the global existence of the solutions and with this the existence of the flow of (17.1), the following theorems are useful.

1. Criterion of Wintner and Conti If $M = \mathbb{R}^n$ in (17.1) and there exists a continuous function $\omega\colon [0, +\infty) \to [1, +\infty)$, such that $\|f(x)\| \le \omega(\|x\|)$ for all $x \in \mathbb{R}^n$ and $\displaystyle\int_0^{+\infty} \frac{1}{\omega(r)}\, dr = +\infty$ holds, then every solution of (17.1) can be extended onto the whole of $\mathbb{R}_+$.

■ For example, the following functions satisfy the criterion of Wintner and Conti: $\omega(r) = Cr + 1$ or $\omega(r) = C\, r |\ln r| + 1$, where $C > 0$ is a constant.

2. Extension Principle If a solution of (17.1) stays bounded as time increases, then it can be extended to the whole of $\mathbb{R}_+$.

Assumption: In the following discussion the existence of the flow $\{\varphi^t\}_{t \in \mathrm{R}}$ of (17.1) is always assumed.

2. Phase Portrait

a) If $\varphi(t)$ is a solution of (17.1), then the function $\varphi(t + c)$ with an arbitrary constant c is also a solution.

b) Two arbitrary orbits of (17.1) have no common point or they coincide. Hence, the phase space of (17.1) is decomposed into disjoint orbits. The decomposition of the phase space into disjoint orbits is called a *phase portrait*.

c) Every orbit, different from a steady state, is a regular smooth curve, which can be closed or not closed.

3. Liouville's Theorem

Let $\{\varphi^t\}_{t\in\mathbb{R}}$ be the flow of (17.1), $D \subset M \subset \mathbb{R}^n$ be an arbitrary bounded and measurable set, $D_t := \varphi^t(D)$ and $V_t := \text{vol}(D_t)$ be the n-dimensional volume of D_t **(Fig. 17.2a)**. Then the relation $\frac{d}{dt} V_t = \int_{D_t} \text{div} f(x)\, dx$ holds for arbitrary $t \in \mathbb{R}$. For $n = 3$, Liouville's theorem states:

$$\frac{d}{dt} V_t = \iiint_{D_t} \text{div} f(x_1, x_2, x_3)\, dx_1\, dx_2\, dx_3. \quad (17.12)$$

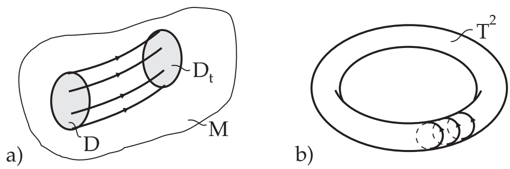


Figure 17.2

Corollary: If $\text{div} f(x) < 0$ in M holds for (17.1), then the flow of (17.1) is volume contracting. If $\text{div} f(x) \equiv 0$ in M holds, then the flow of (17.1) is volume preserving.

■ **A:** For the Lorenz system (17.2), $\text{div} f(x,y,z) \equiv -(\sigma+1+b)$. Since $\sigma > 0$ and $b > 0$, $\text{div} f(x,y,z) < 0$ holds. With Liouville's theorem, $\frac{d}{dt} V_t = \iiint_{D_t} -(\sigma+1+b)\, dx_1\, dx_2\, dx_3 = -(\sigma+1+b)\, V_t$ obviously holds for any arbitrary bounded and measurable set $D \subset \mathbb{R}^3$. The solution of the linear differential equation $\dot{V}_t = -(\sigma+1+b)\, V_t$ is $V_t = V_0 \cdot e^{-(\sigma+1+b)t}$, so that $V_t \to 0$ follows for $t \to +\infty$.

■ **B:** Let $U \subset \mathbb{R}^n \times \mathbb{R}^n$ be an open subset and $H\colon U \to \mathbb{R}$ a C^2-function. Then, $\dot{x}_i = \frac{\partial H}{\partial y_i}(x,y)$, $\dot{y}_i = -\frac{\partial H}{\partial x_i}(x,y)$ $\quad (i = 1, 2, \ldots, n)$ is called a *Hamiltonian differential equation*. The function H is called the *Hamiltonian* of the system. If f denotes the right-hand side of this differential equation, then obviously $\text{div} f(x,y) = \sum_{i=1}^{n} \left[\frac{\partial^2 H}{\partial x_i \partial y_i}(x,y) - \frac{\partial^2 H}{\partial y_i \partial x_i}(x,y) \right] \equiv 0$. Hence, the Hamiltonian differential equations are volume preserving.

17.1.2.2 Linear Differential Equations

1. General Statements

Let $A(t) = [a_{ij}(t)]_{i,j=1}^n$ be a matrix function on $\mathbb{R}$, where every component $a_{ij}\colon \mathbb{R} \to \mathbb{R}$ is a continuous function, and let $b\colon \mathbb{R} \to \mathbb{R}^n$ be a continuous vector function on $\mathbb{R}$. Then

$$\dot{x} = A(t)x + b(t) \quad (17.13a)$$

is called an *inhomogeneous linear first-order differential equation* in $\mathbb{R}^n$, and

$$\dot{x} = A(t)x \quad (17.13b)$$

is the corresponding *homogeneous linear first-order differential equation*.

1. Fundamental Theorem for Homogeneous Linear Differential Equations Every solution of (17.13a) exists on the whole of $\mathbb{R}$. The set of all solutions of (17.13b) forms an n-dimensional vector subspace L_H of the C^1-smooth vector functions over $\mathbb{R}$.

2. Fundamental Theorem for Inhomogeneous Linear Differential Equations The set of all solutions L_I of (17.13a) is an n-dimensional affine vector subspace of the C^1-smooth vector functions over $\mathbb{R}$ of the form $L_I = \varphi_0 + L_H$, where φ_0 is an arbitrary solution of (17.13a).

Let $\varphi_1, \ldots, \varphi_n$ be arbitrary solutions of (17.13b) and $\Phi = [\varphi_1, \ldots, \varphi_n]$ the corresponding *solution matrix*. Then Φ satisfies the *matrix differential equation* $\dot{Z}(t) = A(t)Z(t)$ on $\mathbb{R}$, where $Z \in \mathbb{R}^{n \cdot n}$. If the solutions $\varphi_1, \ldots, \varphi_n$ form a basis of L_H, then $\Phi = [\varphi_1, \ldots, \varphi_n]$ is called the *fundamental matrix* of (17.13b). $W(t) = \det \Phi(t)$ is the *Wronskian determinant* with respect to the solution matrix Φ of (17.13b). The *formula of Liouville* states that:

$$\dot{W}(t) = \operatorname{rank} A(t)\, W(t) \quad (t \in \mathbb{R}). \tag{17.13c}$$

For a solution matrix, either $W(t) \equiv 0$ on $\mathbb{R}$ or $W(t) \neq 0$ for all $t \in \mathbb{R}$. The system $\varphi_1, \ldots, \varphi_n$ is a basis of L_H, if and only if $\det[\varphi_1(t), \ldots, \varphi_n(t)] \neq 0$ for a t (and so for all t).

3. Theorem (Variation of Constants Formula) Let Φ be an arbitrary fundamental matrix of (17.13b). Then the solution φ of (17.13a) with initial point p at time $t = \tau$ can be represented in the form

$$\varphi(t) = \Phi(t)\Phi(\tau)^{-1}\, p + \int_{\tau}^{t} \Phi(t)\Phi(s)^{-1}\, b(s)\, ds \quad (t \in \mathbb{R}). \tag{17.13d}$$

2. Autonomous Linear Differential Equations

Consider the differential equation

$$\dot{x} = A\,x, \tag{17.14}$$

where A is a constant matrix of type (n, n). The *operator norm* (see also 12.5.1.1, p. 677) of a matrix A is given by $\|A\| = \max\{\|A\,x\|,\ x \in \mathbb{R}^n,\ \|x\| \leq 1\}$, where for the vectors of $\mathbb{R}^n$ the Euclidean norm is again considered.

Let A and B be two arbitrary matrices of type (n, n). Then

a) $\|A + B\| \leq \|A\| + \|B\|$, **b)** $\|\lambda A\| = |\lambda|\ \|A\|\ (\lambda \in \mathbb{R})$,

c) $\|Ax\| \leq \|A\|\|x\|\ x \in \mathbb{R}^n)$, **d)** $\|AB\| \leq \|A\|\|B\|$,

e) $\|A\| = \sqrt{\lambda_{\max}}$, where $\lambda_{\max}$ is the greatest eigenvalue of $A^T A$.

The fundamental matrix with initial value E_n at time $t = 0$ of (17.14) is the *matrix-exponential function* (see 5.3.6.4,**4.**, p. 356)

$$e^{At} = E_n + \frac{At}{1!} + \frac{A^2t^2}{2!} + \cdots = \sum_{i=0}^{\infty} \frac{A^i t^i}{i!} \tag{17.15}$$

with the following properties:

a) The series of e^{At} is uniformly convergent with respect to t on an arbitrary compact time interval and absolutely convergent for every fixed t;

b) $\|e^{At}\| \leq e^{\|A\|t} \quad (t \geq 0)$;

c) $\dfrac{d}{dt}(e^{At}) = (e^{At})^{\cdot} = Ae^{At} = e^{At}A \quad (t \in \mathbb{R})$;

d) $e^{(t+s)A} = e^{tA}\, e^{sA} \quad (s, t \in \mathbb{R})$;

e) e^{At} is regular for all t and $(e^{At})^{-1} = e^{-At}$;

f) if A and B are commutative matrices of type (n, n), i.e., $AB = BA$ holds, then $B\, e^A = e^A\, B$ and $e^{A+B} = e^A\, e^B$;

g) if A and B are matrices of type (n, n) and B is regular, then $e^{BAB^{-1}} = B\, e^A\, B^{-1}$.

3. Linear Differential Equations with Periodic Coefficients

Considered is the homogeneous linear differential equation (17.13b), where $A(t) = [a_{ij}(t)]_{i,j=1}^{n}$ is a T-*periodic matrix function*, i.e., $a_{ij}(t) = a_{ij}(t+T) \quad (\forall t \in \mathbb{R},\ i,j = 1,\ldots,n)$. In this case (17.13b) is called a *linear T-periodic differential equation*. Then every fundamental matrix Φ of (17.13b) can be written in the form $\Phi(t) = G(t)e^{tR}$, where $G(t)$ is a smooth, regular T-periodic matrix function and R is a constant matrix of type (n,n) (*Floquet's theorem*).

Let $\Phi(t)$ be the fundamental matrix of the T-periodic differential equation (17.13b), normed at $t = 0$, i.e., $\Phi(0) = E_n$, and let $\Phi(t) = G(t)e^{tR}$ be a representation of it according to Floquet's theorem. The matrix $\Phi(T) = e^{RT}$ is called the *monodromy matrix* of (17.13b); the eigenvalues ρ_j of $\Phi(T)$ are the *multipliers* of (17.13b). A number $\rho \in \mathbf{C}$ is a multiplier of (17.13b) if and only if there exists a solution $\varphi \not\equiv 0$ of (17.13b) such that $\varphi(t+T) = \rho\,\varphi(t) \quad (t \in \mathbb{R})$ holds.

17.1.2.3 Stability Theory

1. Lyapunov Stability and Orbital Stability

Consider the non-autonomous differential equation (17.11). The solution $\varphi(t,t_0,x_0)$ of (17.11) is said to be *stable in the sense of Lyapunov* if:

$$\forall t_1 \ge t_0\ \forall \varepsilon > 0\ \ \exists \delta = \delta(\varepsilon, t_1) \quad \left.\begin{matrix}\forall x_1 \in M \\ \|x_1 - \varphi(t_1,t_0,x_0)\| < \delta\end{matrix}\right\}: \|\varphi(t,t_1,x_1) - \varphi(t,t_0,x_0)\| < \varepsilon \quad \forall t \ge t_1. \tag{17.16a}$$

The solution $\varphi(t,t_0,x_0)$ is called *asymptotically stable in the sense of Lyapunov*, if it is stable and:

$$\forall t_1 \ge t_0\ \ \exists \Delta = \Delta(t_1) \quad \left.\begin{matrix}\forall x_1 \in M \\ \|x_1 - \varphi(t_1,t_0,x_0)\| < \Delta\end{matrix}\right\}: \|\varphi(t,t_1,x_1) - \varphi(t,t_0,x_0)\| \to 0 \quad \text{for } t \to +\infty. \tag{17.16b}$$

For the autonomous differential equation (17.1), there are other important notions of stability besides the Lyapunov stability. The solution $\varphi(t,x_0)$ of (17.1) is called *orbitally stable* (*asymptotically orbitally stable*), if the orbit $\gamma(x_0) = \{\varphi(t,x_0),\ t \in \mathbb{R}\}$ is stable (asymptotically stable) as an invariant set. A solution of (17.1) which represents an equilibrium point is Lyapunov stable exactly if it is orbitally stable. The two types of stability can be different for periodic solutions of (17.1).

■ Let a flow be given in $\mathbb{R}^3$, whose invariant set is the torus T^2. Locally, let the flow be described in a rectangular coordinate system by $\dot{\Theta}_1 = 0,\ \dot{\Theta}_2 = f_2(\Theta_1)$, where $f_2\colon \mathbb{R} \to \mathbb{R}$ is a 2π periodic smooth function, for which:

$$\forall \Theta_1 \in \mathbb{R} \quad \exists\, U_{\Theta_1} \text{ (neighborhood of } \Theta_1) \quad \left.\begin{matrix}\forall \delta_1, \delta_2 \in U_{\Theta_1} \\ \delta_1 \ne \delta_2\end{matrix}\right\}: f_2(\delta_1) \ne f_2(\delta_2).$$

An arbitrary solution satisfying the initial conditions $(\Theta_1(0),\ \Theta_2(0))$ can be given on the torus by

$$\Theta_1(t) \equiv \Theta_1(0), \quad \Theta_2(t) = \Theta_2(0) + f_2(\Theta_1(0)t \quad (t \in \mathbb{R}).$$

From this representation it can be seen that every solution is orbitally stable but not Lyapunov stable **(Fig. 17.2b)**.

2. Asymptotical Stability Theorem of Lyapunov

A scalar-valued function V is called *positive definite* in a neighborhood U of a point $p \in M \subset \mathbb{R}^n$, if:

1. $V\colon U \subset M \to \mathbb{R}$ is continuous.

2. $V(x) > 0$ for all $x \in U \setminus \{p\}$ and $V(p) = 0$.

Let $U \subset M$ be an open subset and $V : U \to \mathbb{R}$ a continuous function. The function V is called a *Lyapunov function* of (17.1) in U, if $V(\varphi(t))$ does not increase while for the solution $\varphi(t) \in U$ holds. Let $V : U \to \mathbb{R}$ be a Lyapunov function of (17.1) and let V be positive definite in a neighborhood U of p. Then p is stable. If the condition $V(\varphi(t,x_0)) =$ constant $(t \ge 0)$ always yields $\varphi(t,x_0) \equiv p$ for

a solution φ of (17.1) with $\varphi(t,x) \in U \quad (t \geq 0)$, i.e., if the Lyapunov function is constant along a complete trajectory, then this trajectory can only be an equilibrium point, and the equilibrium point p is also asymptotically stable.

■ The point $(0,0)$ is a steady point of the *planar differential equation* $\dot{x} = y,\ \dot{y} = -x - x^2 y$. The function $V(x,y) = x^2 + y^2$ is positive definite in every neighborhood of $(0,0)$ and for its derivative $\frac{d}{dt}V(x(t),y(t)) = -2x(t)^2 y(t)^2 < 0$ holds along an arbitrary solution for $x(t)y(t) \neq 0$. Hence, $(0,0)$ is asymptotically stable.

3. Classification and Stability of Steady States

Let x_0 be an equilibrium point of (17.1). In the neighborhood of x_0 the local behavior of the orbits of (17.1) can be described under certain assumptions by the *variational equation* $\dot{y} = D\,f(x_0)y$, where $D\,f(x_0)$ is the Jacobian matrix of f in x_0. If $D\,f(x_0)$ does not have an eigenvalue λ_j with $\mathrm{Re}\,\lambda_j = 0$, then the equilibrium point x_0 is called *hyperbolic*. The hyperbolic equilibrium point x_0 is of type (m,k) if $D\,f(x_0)$ has exactly m eigenvalues with negative real parts and $k = n - 1 - m$ eigenvalues with positive real parts. The hyperbolic equilibrium point of type (m,k) is called a *sink*, if $m = n$, a *source*, if $k = n$, and a *saddle* point, if $m \neq 0$ and $k \neq 0$ **(Fig. 17.3)**. A sink is asymptotically stable; sources and saddles are unstable (*theorem on stability in the first approximation*). Within the three topological basic types of hyperbolic equilibrium points (sink, source, saddle point) further algebraic distinctions can be made. A sink (source) is called a *stable node* (*unstable node*) if every eigenvalue of the Jacobian matrix is real, and a *stable focus* (*unstable focus*) if there are eigenvalues with non-vanishing imaginary parts. For $n = 3$ a classification of saddle points is obtained as saddle nodes and saddle foci.

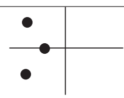
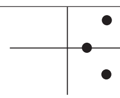
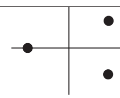
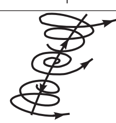
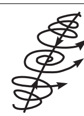
Type of equilibrium point	**Sink**	**Source**	**Saddle point**
Eigenvalues of the Jacobian matrix			
Phase portrait			

Figure 17.3

4. Stability of Periodic Orbits

Let $\varphi(t,x_0)$ be a T-periodic solution of (17.1) and $\gamma(x_0) = \{\varphi(t,x_0),\ t \in [0,T]\}$ its orbit. Under certain assumptions, the phase portrait in a neighborhood of $\gamma(x_0)$ can be described by the *variational equation* $\dot{y} = D\,f(\varphi(t,x_0))\,y$. Since $A(t) = D\,f(\varphi(t,x_0))$ is a T-periodic continuous matrix function of type (n,n), it follows from the Floquet theorem (see p. 863) that the fundamental matrix $\Phi_{x_0}(t)$ of the variational equation can be written in the form $\Phi_{x_0}(t) = G(t)e^{Rt}$, where G is a T-periodic regular smooth matrix function with $G(0) = E_n$, and R represents a constant matrix of type (n,n) which is not uniquely given. The matrix $\Phi_{x_0}(T) = e^{RT}$ is called the *monodromy matrix of the periodic orbit* $\gamma(x_0)$, and the eigenvalues $\rho_1, \ldots, \rho_n$ of e^{RT} are called *multipliers of the periodic orbit* $\gamma(x_0)$. If the orbit $\gamma(x_0)$ is represented by another solution $\varphi(t,x_1)$, i.e., if $\gamma(x_0) = \gamma(x_1)$, then the multipliers of $\gamma(x_0)$ and $\gamma(x_1)$ coincide. One of the multipliers of a periodic orbit is always equal to one (*Andronov–Witt theorem*). Let $\rho_1, \ldots, \rho_{n-1}, \rho_n = 1$ be the multipliers of the periodic orbit $\gamma(x_0)$ and let $\Phi_{x_0}(T)$ be the monodromy matrix of $\gamma(x_0)$. Then

$$\sum_{j=1}^{n} \rho_j = \mathrm{Tr}\,\Phi_{x_0}(T) \text{ and } \prod_{j=1}^{n} \rho_j = \det \Phi_{x_0}(T) = \exp\left(\int_0^T \mathrm{Tr} D\,f(\varphi(t,x_0))dt\right)$$

$$= \exp\left(\int_0^T \operatorname{div} f(\varphi(t,x_0))\,dt\right). \qquad (17.17)$$

Hence, if $n = 2$, then $\rho_2 = 1$ and $\rho_1 = \exp\left(\int_0^T \operatorname{div} f(\varphi(t,x_0))\,dt\right)$.

■ Let $\varphi(t,(1,0)) = (\cos t, \sin t)$ be a 2π-periodic solution of (17.9a). The matrix $A(t)$ of the variational equation with respect to this solution is

$$A(t) = D\,f(\varphi(t,(1,0))) = \begin{pmatrix} -2\cos^2 t & -1-\sin 2t \\ 1-\sin 2t & -2\sin^2 t \end{pmatrix}.$$

The fundamental matrix $\Phi_{(1,0)}(t)$ normed at $t = 0$ is given by

$$\Phi_{(1,0)}(t) = \begin{pmatrix} e^{-2t}\cos t & -\sin t \\ e^{-2t}\sin t & \cos t \end{pmatrix} = \begin{pmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{pmatrix} \begin{pmatrix} e^{-2t} & 0 \\ 0 & 1 \end{pmatrix},$$

where the last product is a Floquet representation of $\Phi_{(1,0)}(t)$. Thus, $\rho_1 = e^{-4\pi}$ and $\rho_2 = 1$. The multipliers can be determined without the Floquet representation. For system (17.9a) $\operatorname{div} f(x,y) = 2-4x^2-4y^2$ holds, and hence $\operatorname{div} f(\cos t, \sin t) \equiv -2$. According to the formula above, $\rho_1 = \exp\left(\int_0^{2\pi} -2dt\right) = \exp(-4\pi)$.

5. Classification of Periodic Orbits

If the periodic orbit γ of (17.1) has no further multiplier on the complex unit circle besides $\rho_n = 1$, then γ is called *hyperbolic*. The hyperbolic periodic orbit is of *type* (m,k) if there are m multipliers inside and $k = n-1-m$ multipliers outside the unit circle. If $m > 0$ and $k > 0$, then the periodic orbit of type (m,k) is called a *saddle point*.

According to the Andronov-Witt theorem, a hyperbolic periodic orbit γ of (17.1) of type $(n-1,0)$ is asymptotically stable. Hyperbolic periodic orbits of type (m,k) with $k > 0$ are unstable.

■ **A:** A periodic orbit $\gamma = \{\varphi(t), t \in [0,T]\}$ in the plane with multipliers ρ_1 and $\rho_2 = 1$ is asymptotically stable if $|\rho_1| < 1$, i.e., if $\int_0^T \operatorname{div} f(\varphi(t))\,dt < 0$.

■ **B:** If there is a further multiplier besides $\rho_n = 1$ on the complex unit circle, then the Andronov–Witt theorem cannot be applied. The information about the multipliers is not sufficient for the stability analysis of the periodic orbit.

■ **C:** As an example, let the planar system $\dot{x} = -y + x\,f(x^2+y^2)$, $\quad \dot{y} = x + y\,f(x^2+y^2)$ be given by the smooth function $f\colon (0,+\infty) \to \mathbb{R}$, which additionally satisfies the properties $f(1) = f'(1) = 0$ and $f(r)(r-1) < 0$ for all $r \neq 1, r > 0$. Obviously, $\varphi(t) = (\cos t, \sin t)$ is a 2π-periodic solution of the system and

$\Phi_{(1,0)}(t) = \begin{pmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ is the Floquet representation of the fundamental matrix. It follows that $\rho_1 = \rho_2 = 1$. The use of polar coordinates results in the system $\dot{r} = r\,f(r^2)$, $\dot{\vartheta} = 1$. This representation yields that the periodic orbit $\gamma((1,0))$ is asymptotically stable.

6. Properties of Limit Sets, Limit Cycles

The α- and ω-limit sets defined in 17.1.1.2, p. 859, have with respect to the flow of the differential equation (17.1) with $M \subset \mathbb{R}^n$ the following properties. Let $x \in M$ be an arbitrary point. Then:

a) The sets $\alpha(x)$ and $\omega(x)$ are closed.

b) If $\gamma^+(x)$ (respectively $\gamma^-(x)$) is bounded, then $\omega(x) \neq \emptyset$ (respectively $\alpha(x) \neq \emptyset$) holds. Furthermore, $\omega(x)$ (respectively $\alpha(x)$) are in this case invariant under the flow (17.1) and connected.

■ If for instance, $\gamma^+(x)$ is not bounded, then $\omega(x)$ is not necessarily connected **(Fig. 17.4a)**.

For a planar autonomous differential equation (17.1), (i.e., $M \subset \mathbb{R}^2$) the Poincaré–Bendixson theorem is valid.

Poincaré–Bendixson Theorem: Let $\varphi(\cdot,p)$ be a non-periodic solution of (17.1), for which $\gamma^+(p)$ is

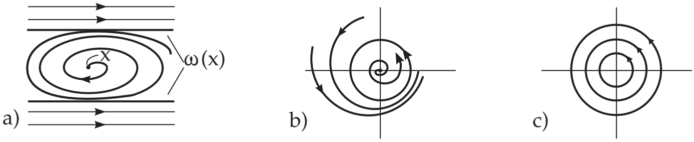


Figure 17.4

bounded. If $\omega(p)$ contains no equilibrium point of (17.1), then $\omega(p)$ is a periodic orbit of (17.1). Hence, for autonomous differential equations in the plane, attractors more complicated than an equilibrium point or a periodic orbit are not possible.

A periodic orbit γ of (17.1) is called a *limit cycle*, if there exists an $x \notin \gamma$ such that either $\gamma \subset \omega(x)$ or $\gamma \subset \alpha(x)$ holds. A limit cycle is called a *stable limit cycle* if there exists a neighborhood U of γ such that $\gamma = \omega(x)$ holds for all $x \in U$, and an *unstable limit cycle* if there exists a neighborhood U of γ such that $\gamma = \alpha(x)$ holds for all $x \in U$.

■ **A:** For the flow of (17.9a), the property $\gamma = \omega(p)$ for all $p \neq (0,0)$ is valid for the periodic orbit $\gamma = \{(\cos t, \sin t),\ \ t \in [0, 2\pi)\}$. Hence, $U = \mathbb{R}^2 \backslash \{(0,0)\}$ is a neighborhood of γ such that with it, γ is a stable limit cycle **(Fig. 17.4b)**.

■ **B:** In contrast, for the linear differential equation $\dot{x} = -y$, $\dot{y} = x$, the orbit $\gamma = \{(\cos t,\ \sin t),\ \ t \in [0, 2\pi]\}$ is a periodic orbit, but not a limit cycle **(Fig. 17.4c)**.

7. *m*-Dimensional Embedded Tori as Invariant Sets

A differential equation (17.1) can have an m-dimensional torus as an invariant set. An *m-dimensional torus T^m embedded* into the phase space $M \subset \mathbb{R}^n$ is defined by a differentiable mapping $g\colon \mathbb{R}^m \to \mathbb{R}^n$, which is supposed to be 2π-periodic in every coordinate Θ_i as a function $(\Theta_1, \ldots, \Theta_m) \mapsto g(\Theta_1, \ldots, \Theta_m)$.

■ In simple cases, the motion of the system (17.1) on the torus can be described in a rightangular coordinate system by the differential equations $\dot{\Theta}_i = \omega_i \quad (i = 1, 2, \ldots, m)$. The solution of this system with initial values $(\Theta_1(0), \ldots, \Theta_m(0))$ at time $t = 0$ is $\Theta_i(t) = \omega_i t + \Theta_i(0) \quad (i = 1, 2, \ldots, m; t \in \mathbb{R})$.

A continuous function $f\colon \mathbb{R} \to \mathbb{R}^n$ is called *quasiperiodic* if f has a representation of the form $f(t) = g(\omega_1 t, \omega_2 t, \ldots, \omega_n t)$, where g is also a differentiable function as above, which is 2π-periodic in every component, and the frequencies ω_i are *incommensurable*, i.e., there are no such integers n_i with $\sum\limits_{i=1}^{m} n_i^2 > 0$ for which $n_1\omega_1 + \cdots + n_m\omega_m = 0$ holds.

17.1.2.4 Invariant Manifolds

1. Definition, Separatrix Surfaces

Let γ be a hyperbolic equilibrium point or a hyperbolic periodic orbit of (17.1). The *stable manifold* $W^s(\gamma)$ (respectively *unstable manifold* $W^u(\gamma)$) of γ is the set of all points of the phase space such that the orbits tending to γ as $t \to +\infty$ (respectively $t \to -\infty$) pass through these points:

$$W^s(\gamma) = \{x \in M\colon\ \omega(x) = \gamma\} \ \text{ and } \ W^u(\gamma) = \{x \in M\colon\ \alpha(x) = \gamma\}. \tag{17.18}$$

Stable and unstable manifolds are also called *separatrix surfaces*.

■ In the plane, the differential equation

$$\dot{x} = -x, \quad \dot{y} = y + x^2 \tag{17.19a}$$

is considered. The solution of (17.19a) with initial state (x_0, y_0) at time $t = 0$ is explicitly given by

$$\varphi(t, x_0, y_0) = (e^{-t}x_0, e^t y_0 + \frac{x_0^2}{3}\,(e^t - e^{-2t}))\,. \tag{17.19b}$$

For the stable and unstable manifolds of the equilibrium point $(0,0)$ of (17.19a) one gets:

$$W^s((0,0)) = \{(x_0, y_0)\colon \lim_{t\to+\infty} \varphi(t, x_0, y_0) = (0,0)\} = \{(x_0, y_0)\colon y_0 + \frac{x_0^2}{3} = 0\},$$

$W^u((0,0)) = \{(x_0, y_0)\colon \lim_{t\to-\infty} \varphi(t, x_0, y_0) = (0,0)\} = \{(x_0, y_0)\colon x_0 = 0,\ y_0 \in \mathbb{R}\}$ **(Fig. 17.5a)**.

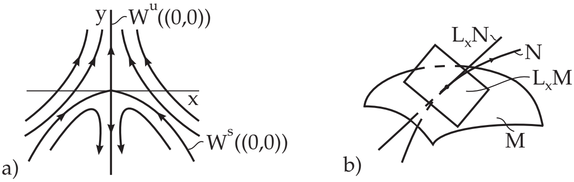


Figure 17.5

Let M and N be two smooth surfaces in $\mathbb{R}^n$, and let L_xM and L_xN be the corresponding tangent planes to M and N through x. The surfaces M and N are *transversal* to each other if for all $x \in M \cap N$ the following relation holds:

$$\dim L_xM + \dim L_xN - n = \dim\,(L_xM \cap L_xN).$$

■ For the section represented in **Fig. 17.5b** one can see that $\dim L_xM = 2$, $\dim L_xN = 1$ and $\dim(L_xM \cap L_xN) = 0$ holds. Hence, the section represented in **Fig. 17.5b** is transversal.

2. Theorem of Hadamard and Perron

Important properties of separatrix surfaces are given by the *Theorem of Hadamard and Perron*:
Let γ be a hyperbolic equilibrium point or a hyperbolic periodic orbit of (17.1).

a) The manifolds $W^s(\gamma)$ and $W^u(\gamma)$ are generalized C^r-surfaces, which locally look like C^r-smooth elementary surfaces. Every orbit of (17.1), which does not tend to γ for $t \to +\infty$ or $t \to -\infty$, respectively, leaves a sufficiently small neighborhood of γ for $t \to +\infty$ or $t \to -\infty$, respectively.

b) If $\gamma = x_0$ is an equilibrium point of type (m,k), then $W^s(x_0)$ and $W^u(x_0)$ are surfaces of dimension m and k, respectively. The surfaces $W^s(x_0)$ and $W^u(x_0)$ are tangent at x_0 to the *stable vector subspace*

$$E^s = \{y \in \mathbb{R}^n\colon e^{D\,f(x_0)t}y \to 0 \ \text{ for }\ t \to +\infty\} \ \text{ of equation }\ \dot{y} = D\,f(x_0)y \tag{17.20a}$$

and the *unstable vector subspace*

$$E^u = \{y \in \mathbb{R}^n\colon e^{D\,f(x_0)t}y \to 0 \ \text{ for }\ t \to -\infty\} \ \text{ of equation }\ \dot{y} = D\,f(x_0)y, \ \text{ respectively.} \tag{17.20b}$$

c) If γ is a hyperbolic periodic orbit of type (m,k), then $W^s(\gamma)$ and $W^u(\gamma)$ are surfaces of dimension $m+1$ and $k+1$, respectively, and they intersect each other transversally along γ **(Fig. 17.6a)**.

■ **A:** To determine a local stable manifold through the steady state $(0,0)$ of the differential equation (17.19a) here it is supposed that $W^s_{loc}((0,0))$ has the following form:
$W^s_{loc}((0,0)) = \{(x,y)\colon y = h(x),\ |x| < \Delta,\ h\colon (-\Delta, \Delta) \to \mathbb{R}\ \text{ differentiable}\}$.
Let $(x(t), y(t))$ be a solution of (17.19a) lying in $W^s_{loc}((0,0))$. Based on the invariance, for times s near to t $y(s) = h(x(s))$ holds. By differentiation and representation of $\dot{x}$ and $\dot{y}$ from the system (17.19a) one gets the initial value problem $h'(x)\,(-x) = h(x) + x^2$, $h(0) = 0$ for the unknown function $h(x)$. If looking for the solution in the form of a series expansion $h(x) = \dfrac{a_2}{2}x^2 + \dfrac{a_3}{3!}x^3 + \cdots$, where $h'(0) = 0$ is taken under consideration, then one gets by comparing the coefficients $a_2 = -\dfrac{2}{3}$ and $a_k = 0$ for $k \ge 3$.

■ **B:** For the system

$$\dot{x} = -y + x(1 - x^2 - y^2),\ \dot{y} = x + y(1 - x^2 - y^2),\ \dot{z} = \alpha z \tag{17.21}$$

with a parameter $\alpha > 0$, the orbit $\gamma = \{(\cos t, \sin t, 0),\ t \in [0, 2\pi]\}$ is a periodic orbit with multipliers $\rho_1 = e^{-4\pi}$, $\rho_2 = e^{\alpha 2\pi}$ and $\rho_3 = 1$.
In cylindrical coordinates $x = r\cos\vartheta,\ y = r\sin\vartheta, z = z$, with initial values (r_0, ϑ_0, z_0) at time $t = 0$,

the solution of (17.21) has the representation $(r(t, r_0), \vartheta(t, \vartheta_0), e^{\alpha t} z_0)$, where $r(t, r_0)$ and $\vartheta(t, \vartheta_0)$ is the solution of (17.9a) in polar coordinates. Consequently,

$$W^s(\gamma) = \{(x, y, z) \colon z = 0\} \setminus \{(0, 0, 0)\} \quad \text{and} \quad W^u(\gamma) = \{(x, y, z) \colon x^2 + y^2 = 1\} \quad \text{(cylinder)}.$$

Both separatrix surfaces are shown in **Fig. 17.6b**.

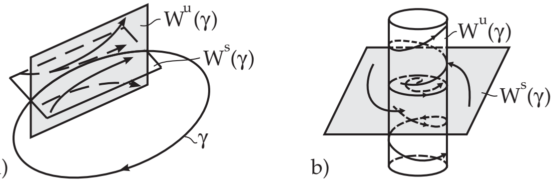


Figure 17.6

3. Local Phase Portraits Near Steady States for $n = 3$

Considered is the differential equation (17.1) with the hyperbolic equilibrium point 0 for $n = 3$. Set $A = D\,f(0)$ and let $\det[\lambda E - A] = \lambda^3 + p\lambda^2 + q\lambda + r$ be the characteristic polynomial of A. With the notation $\delta = p\,q - r$ and $\Delta = -p^2q^2 + 4p^3r + 4q^3 - 18pqr + 27r^2$ (discriminant of the characteristic polynomial), the different equilibrium point types are characterized in **Table 17.1**.

4. Homoclinic and Heteroclinic Orbits

Suppose γ_1 and γ_2 are two hyperbolic equilibrium points or periodic orbits of (17.1). If the separatrix surfaces $W^s(\gamma_1)$ and $W^u(\gamma_2)$ intersect each other, then the intersection consists of complete orbits. For two equilibrium points or periodic orbits, the orbit $\gamma \subset W^s(\gamma_1) \cap W^u(\gamma_2)$ is called *heteroclinic* if $\gamma_1 \neq \gamma_2$ **(Fig. 17.7a)**, and *homoclinic* if $\gamma_1 = \gamma_2$. Homoclinic orbits of equilibrium points are also called *separatrix loops* **(Fig. 17.7b)**.

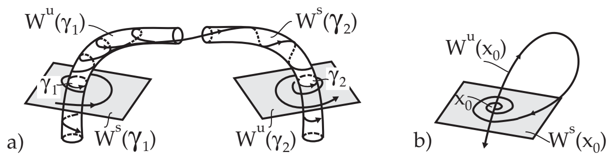


Figure 17.7

■ Consider the Lorenz system (17.2) with fixed parameters $\sigma = 10$, $b = 8/3$ and with variable r. The equilibrium point $(0, 0, 0)$ of (17.2) is a saddle for $1 < r < 13.926\ldots$, which is characterized by a two-dimensional stable manifold W^s and a one-dimensional unstable manifold W^u. If $r = 13.926\ldots$, then there are two separatrix loops at $(0, 0, 0)$, i.e., as $t \to +\infty$ branches of the unstable manifold return (over the stable manifold) to the origin (see [17.9]).

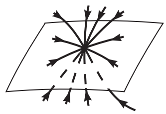
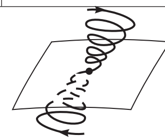
17.1.2.5 Poincaré Mapping

1. Poincaré Mapping for Autonomous Differential Equations

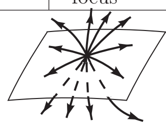
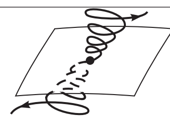
Let $\gamma = \{\varphi(t, x_0),\ t \in [0, T]\}$ be a T-periodic orbit of (17.1) and $\sum$ a $(n-1)$-dimensional smooth hypersurface, which intersects the orbit γ transversally in x_0 **(Fig. 17.8a)**. Then, there is a neighborhood U of x_0 and a smooth function $\tau \colon U \to \mathbb{R}$ such that $\tau(x_0) = T$ and $\varphi(\tau(x), x) \in \sum$ for all $x \in U$. The mapping $P \colon U \cap \sum \to \sum$ with $P(x) = \varphi(\tau(x), x)$ is called the *Poincaré mapping* of γ at x_0. If the right-hand side f of (17.1) is r times continuously differentiable, then P is also r times continuously differentiable . The eigenvalues of the Jacobi matrix $DP(x_0)$ are the multipliers $\rho_1, \ldots, \rho_{n-1}$ of the periodic orbit. They do not depend on the choice of x_0 on γ and on the choice of the transversal surface.

Table 17.1 Steady state types in three-dimensional phase spaces

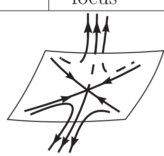
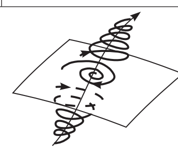
Parameter domain	$\boldsymbol{\Delta}$	**Type of equilibrium point**	**Roots of the characteristic polynomial**	**Dimension of $\boldsymbol{W^s}$ and $\boldsymbol{W^u}$**
$\delta > 0;\ r > 0,$ $q > 0$	$\Delta < 0$	stable node	$\mathrm{Im}\lambda_j = 0$ $\lambda_j < 0,\ j = 1,2,3$	$\dim W^s = 3,\ \dim W^u = 0$
	$\Delta > 0$	stable focus	$\mathrm{Re}\lambda_{1,2} < 0$ $\lambda_3 < 0$	

$\Delta < 0$:  $\Delta > 0$: 

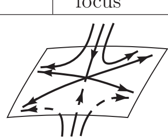
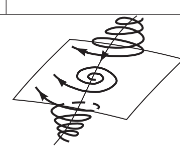
Parameter domain	$\boldsymbol{\Delta}$	**Type of equilibrium point**	**Roots of the characteristic polynomial**	**Dimension of $\boldsymbol{W^s}$ and $\boldsymbol{W^u}$**
$\delta < 0;\ r < 0,$ $q > 0$	$\Delta < 0$	unstable node	$\mathrm{Im}\lambda_j = 0$ $\lambda_j > 0,\ j = 1,2,3$	$\dim W^s = 0,\ \dim W^u = 3$
	$\Delta > 0$	unstable focus	$\mathrm{Re}\lambda_{1,2} > 0$ $\lambda_3 > 0$	

$\Delta < 0$:  $\Delta > 0$: 

Parameter domain	$\boldsymbol{\Delta}$	**Type of equilibrium point**	**Roots of the characteristic polynomial**	**Dimension of $\boldsymbol{W^s}$ and $\boldsymbol{W^u}$**
$\delta > 0;\ r < 0,$ $q \le 0$ or $r < 0,\ q > 0$	$\Delta < 0$	saddle node	$\mathrm{Im}\lambda_j = 0$ $\lambda_{1,2} < 0,\ \lambda_3 > 0$	$\dim W^s = 2,\ \dim W^u = 1$
	$\Delta > 0$	saddle focus	$\mathrm{Re}\lambda_{1,2} < 0$ $\lambda_3 > 0$	

$\Delta < 0$:  $\Delta > 0$: 

Parameter domain	$\boldsymbol{\Delta}$	**Type of equilibrium point**	**Roots of the characteristic polynomial**	**Dimension of $\boldsymbol{W^s}$ and $\boldsymbol{W^u}$**
$\delta < 0;\ r > 0,$ $q \le 0$ or $r > 0,\ q > 0$	$\Delta < 0$	saddle node	$\mathrm{Im}\lambda_j = 0$ $\lambda_{1,2} > 0,\ \lambda_3 < 0$	$\dim W^s = 1,\ \dim W^u = 2$
	$\Delta > 0$	saddle focus	$\mathrm{Re}\lambda_{1,2} > 0$ $\lambda_3 < 0$	

$\Delta < 0$:  $\Delta > 0$: 

A system (17.3) in $M = U$ can be connected with the Poincaré mapping, which makes sense until the iterates stay in U. The periodic orbits of (17.1) correspond to the equilibrium points of this discrete system, and the stability of these equilibrium points corresponds to the stability of the periodic orbits of (17.1).

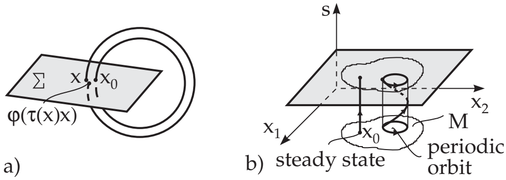


Figure 17.8

■ Considered is for the system (17.9a) the transversal hyperplanes

$$\sum = \{(r, \vartheta)\colon r > 0, \vartheta = \vartheta_0\}$$

in polar coordinate form. For these planes $U = \sum$ can be chosen. Obviously, $\tau(r) = 2\pi$ ($\forall r > 0$) and so

$$P(r) = [1 + (r^{-2} - 1)\, e^{-4\pi}]^{-1/2},$$

where the solution representation of (17.9a) is used. It is also valid that $P(\sum) = \sum$, $P(1) = 1$ and $P'(1) = e^{-4\pi} < 1$.

2. Poincaré Mapping for Non-Autonomous Time-Periodic Differential Equations

A non-autonomous differential equation (17.11), whose right-hand side f has period T with respect to t, i.e., for which $f(t + T, x) = f(t, x)$ ($\forall t \in \mathbb{R}$, $\forall x \in M$) holds, is interpreted as an autonomous differential equation $\dot{x} = f(s, x)$, $\dot{s} = 1$ with cylindrical phase space $M \times \{s \bmod T\}$. Let $s_0 \in \{s \bmod T\}$ be arbitrary. Then, $\sum = M \times \{s_0\}$ is a transversal plane **(Fig. 17.8b)**. The Poincaré mapping is given globally as $P\colon \sum \to \sum$ over $x_0 \longmapsto \varphi(s_0 + T, s_0, x_0)$, where $\varphi(t, s_0, x_0)$ is the solution of (17.11) with the initial state x_0 at time s_0.

17.1.2.6 Topological Equivalence of Differential Equations

1. Definition

Suppose, besides (17.1) with the corresponding flow $\{\varphi^t\}_{t\in\mathbb{R}}$, that a further autonomous differential equation

$$\dot{x} = g(x), \tag{17.22}$$

is given, where $g\colon N \to \mathbb{R}^n$ is a C^r-mapping on the open set $N \subset \mathbb{R}^n$. Of course, the flow $\{\psi^t\}_{t\in\mathbb{R}}$ of (17.22) should exist.

The differential equations (17.1) and (17.22) (or their flows) are called *topologically equivalent* if there exists a *homeomorphism* $h\colon M \to N$ (i.e., h is bijective, h and h^{-1} are continuous), which transforms each orbit of (17.1) to an orbit of (17.22) preserving the orientation, but not necessarily preserving the parametrization. The systems (17.1) and (17.22) are topologically equivalent if there also exists a continuous mapping $\tau\colon \mathbb{R} \times M \to \mathbb{R}$, besides the homeomorphism $h\colon M \to N$, such that τ is strictly monotonically increasing at every fixed $x \in M$, maps $\mathbb{R}$ onto $\mathbb{R}$, with $\tau(0, x) = 0$ for all $x \in M$ and satisfies the relation $h(\varphi^t(x)) = \psi^{\tau(t,x)}(h(x))$ for all $x \in M$ and $t \in \mathbb{R}$.

In the case of topological equivalence, the equilibrium points of (17.1) go over into steady states of (17.22) and periodic orbits of (17.1) go over into periodic orbits of (17.22), where the periods are not necessarily coincident. Hence, if two systems (17.1) and (17.22) are topologically equivalent, then the topological structure of the decomposition of the phase spaces into orbits is the same. If two systems (17.1) and (17.22) are topologically equivalent with the homeomorphism $h\colon M \to N$ and if h preserves the parametrization, i.e., $h(\varphi^t(x)) = \psi^t(h(x))$ holds for every t, x, then (17.1) and (17.22) are called

topologically conjugate.

Topological equivalence or conjugation can also refer to subsets of the phase spaces M and N. Suppose, e.g., (17.1) is defined on $U_1 \subset M$ and (17.22) on $U_2 \subset N$. Then (17.1) *on* U_1 is called *topologically equivalent to* (17.22) *on* U_2 if there exists a homeomorphism $h\colon U_1 \to U_2$ which transforms the intersection of the orbits of (17.1) with U_1 into the intersection of the orbits of (17.22) with U_2 preserving the orientation.

■ **A:** Homeomorphisms for (17.1) and (17.22) are mappings where, e.g., stretching and shrinking of the orbits are allowed; cutting and closing are not.
The flows corresponding to phase portraits of **Fig. 17.9a** and **Fig. 17.9b** are topologically equivalent; the flows shown in **Fig. 17.9a** and **Fig. 17.9c** are not.

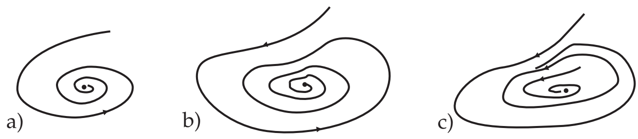


Figure 17.9

■ **B:** Consider the two *linear planar differential equations* (see [17.11])
$\dot{x} = Ax$ and $\dot{x} = Bx$ with $A = \begin{pmatrix} -1 & -3 \\ -3 & -1 \end{pmatrix}$ and $B = \begin{pmatrix} 4 & 0 \\ 0 & -8 \end{pmatrix}$. The phase portraits of these systems close to $(0,0)$ are shown in **Fig. 17.10a** and **Fig. 17.10b**.
The homeomorphism $h: \mathbb{R}^2 \to \mathbb{R}^2$ with $h(x) = Rx$, where $R = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}$, and the function $\tau\colon \mathbb{R} \times \mathbb{R}^2 \to \mathbb{R}$ with $\tau(t,x) = \frac{1}{2}t$ transform the orbits of the first system into the orbits of the second one. Hence, the two systems are topologically equivalent.

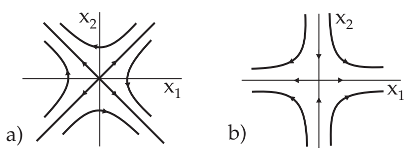


Figure 17.10

2. Theorem of Grobman and Hartman

Let p be a hyperbolic equilibrium point of (17.1). Then, in a neighborhood of p the differential equation (17.1) is topologically equivalent to its linearization $\dot{y} = Df(p)y$.

17.1.3 Discrete Dynamical Systems

17.1.3.1 Steady States, Periodic Orbits and Limit Sets

1. Types of Steady State Points

Let x_0 be an equilibrium point of (17.3) with $M \subset \mathbb{R}^n$. The local behavior of the iteration (17.3) close to x_0 is given, under certain assumptions, by the *variational equation* $y_{t+1} = D\varphi(x_0)y_t,\ t \in \Gamma$. If $D\varphi(x_0)$ has no eigenvalue λ_i with $|\lambda_i| = 1$, then the steady state point x_0, analogously to the differential equation case, is called *hyperbolic*. The hyperbolic equilibrium point x_0 is *of type* (m,k) if $Df(x_0)$ has exactly m eigenvalues inside and $k = n-m$ eigenvalues outside the complex unit circle. The hyperbolic equilibrium point of type (m,k) is called a *sink* for $m = n$, a *source* for $k = n$ and a *saddle point* for

$m > 0$ and $k > 0$. A sink is asymptotically stable; sources and saddles are unstable (*theorem on stability in the first approximation for discrete systems*).

2. Periodic Orbits

Let $\gamma(x_0) = \{\varphi^k(x_0),\ \ k = 0, \cdots, T-1\}$ be a T-periodic orbit ($T \geq 2$) of (17.3). If x_0 is a hyperbolic equilibrium point of the mapping φ^T, then $\gamma(x_0)$ is called *hyperbolic*.
The matrix $D\varphi^T(x_0) = D\varphi(\varphi^{T-1}(x_0)) \cdots D\varphi(x_0)$ is called the *monodromy matrix*; the eigenvalues ρ_i of $D\varphi^T(x_0)$ are the *multipliers* of $\gamma(x_0)$.

If all multipliers ρ_i of $\gamma(x_0)$ have an absolute value less than one, then the periodic orbit $\gamma(x_0)$ is asymptotically stable.

3. Properties of ω-Limit Set

Every ω-limit set $\omega(x)$ of (17.3) with $M = \mathbb{R}^n$ is closed, and $\omega(\varphi(x)) = \omega(x)$. If the semiorbit $\gamma^+(x)$ is bounded, then $\omega(x) \neq \emptyset$ and $\omega(x)$ is invariant under φ. Analogous properties are valid for α-limit sets.

■ Suppose the difference equation $x_{t+1} = -x_t,\ \ t = 0, \pm 1, \cdots$, is given on $\mathbb{R}$ with $\varphi(x) = -x$. Obviously, the relations $\omega(1) = \{1, -1\},\ \ \omega(\varphi(1)) = \omega(-1) = \omega(1)$, and $\varphi(\omega(1)) = \omega(1)$ are satisfied for $x = 1$. It is to be mentioned that $\omega(1)$ is not connected, is different from the case of differential equations.

17.1.3.2 Invariant Manifolds

1. Separatrix Surfaces

Let x_0 be an equilibrium point of (17.3). Then $W^s(x_0) = \{y \in M\colon \varphi^i(y) \to x_0 \text{ for } i \to +\infty\}$ is called a *stable manifold* and $W^u(x_0) = \{y \in M\colon \varphi^i(y) \to x_0 \text{ for } i \to -\infty\}$ an *unstable manifold* of x_0. Stable and unstable manifolds are also called *separatrix surfaces*.

2. Theorem of Hadamard and Perron

The theorem of Hadamard and Perron describes the properties of separatrix surfaces for discrete systems in $M \subset \mathbb{R}^n$:
If x_0 is a hyperbolic equilibrium point of (17.3) of type (m, k), then $W^s(x_0)$ and $W^u(x_0)$ are generalized C^r-smooth surfaces of dimension m and k, respectively, which locally look like C^r-smooth elementary surfaces. The orbits of (17.3), which do not tend to x_0 for $i \to +\infty$ or $i \to -\infty$, leave a sufficiently small neighborhood of x_0 for $i \to +\infty$ or $i \to -\infty$, respectively. The surfaces $W^s(x_0)$ and $W^u(x_0)$ are tangent at x_0 to the *stable vector subspace* $E^s = \{y \in \mathbb{R}^n : [D\varphi(x_0)]^i\ y \to 0 \text{ for } i \to -\infty\}$ of $y_{i+1} = D\varphi(x_0)y_i$ and the *unstable vector subspace* $E^u = \{y \in \mathbb{R}^n : [D\varphi(x_0)]^i y \to 0 \text{ for } i \to -\infty\}$, respectively.

■ Considered is the following time discrete dynamical system from the family of Hénon mappings:

$$x_{i+1} = x_i^2 + y_i - 2,\ \ y_{i+1} = x_i,\ \ i \in \mathbb{Z}. \tag{17.23}$$

Both hyperbolic equilibrium points of (17.23) are $P_1 = (\sqrt{2}, \sqrt{2})$ and $P_2 = (-\sqrt{2}, -\sqrt{2})$.
Determination of the local stable and unstable manifolds of P_1: The variable transformation $x_i = \xi_i + \sqrt{2},\ \ y_i = \eta_i + \sqrt{2}$ transforms system (17.23) into the system $\xi_{i+1} = \xi_i^2 + 2\sqrt{2}\ \xi_i + \eta_i,\ \ \eta_{i+1} = \xi_i$ with the equilibrium point $(0, 0)$. The eigenvectors $a_1 = (\sqrt{2} + \sqrt{3}, 1)$ and $a_2 = (\sqrt{2} - \sqrt{3}, 1)$ of the Jacobian matrix $Df((0, 0))$ correspond to the eigenvalues $\lambda_{1,2} = \sqrt{2} \pm \sqrt{3}$, so $E^s = \{ta_2, t \in \mathbb{R}\}$ and $E^u = \{ta_1, t \in \mathbb{R}\}$. Supposing that $W^u_{loc}((0, 0)) = \{(\xi, \eta)\colon \eta = \beta(\xi),\ \ |\xi| < \Delta, \beta\colon (-\Delta, \Delta) \to \mathbb{R} \text{ differentiable}\}$, one looks for β in the form of a power series $\beta(\xi) = (\sqrt{3} - \sqrt{2})\ \ \xi + k\xi^2 + \cdots$. From $(\xi_i, \eta_i) \in W^u_{loc}((0, 0))$, $(\xi_{i+1}, \eta_{i+1}) \in W^u_{loc}((0, 0))$ follows. This leads to an equation for the coefficients of the decomposition of β, where $k < 0$. The theoretical shape of the stable and unstable manifolds is shown in **Fig. 17.11a** (see [17.12]).

3. Transverse Homoclinic Points

The separatrix surfaces $W^s(x_0)$ and $W^u(x_0)$ of a hyperbolic equilibrium point x_0 of (17.3) can intersect each other. If the intersection $W^s(x_0) \cap W^u(x_0)$ is transversal, then every point $y \in W^s(x_0) \cap W^u(x_0)$ is called a *transversal homoclinic point*.

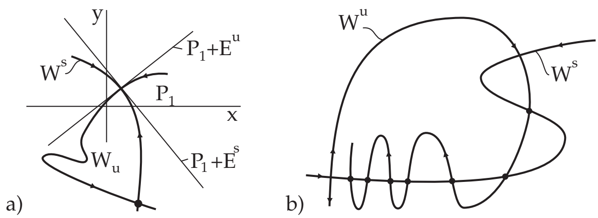


Figure 17.11

Fact: If y is a transversal homoclinic point, then the orbit $\{\varphi^i(y)\}$ of the invertible system (17.3) consists only of transversal homoclinic points **(Fig. 17.11b)**.

17.1.3.3 Topological Conjugation of Discrete Systems

1. Definition

Suppose, besides (17.3), a further discrete system

$$x_{t+1} = \psi(x_t) \tag{17.24}$$

with $\psi\colon N \to N$ is given, where $N \subset \mathbb{R}^n$ is an arbitrary set and ψ is continuous (M and N can be general metric spaces). The discrete systems (17.3) and (17.24) (or the mappings φ and ψ) are called *topologically conjugate* if there exists a homeomorphism $h\colon M \to N$ such that $\varphi = h^{-1} \circ \psi \circ h$. If (17.3) and (17.24) are topologically conjugated, then the homeomorphism h transforms the orbits of (17.3) into orbits of (17.24).

2. Theorem of Grobman and Hartman

If φ in (17.3) is a diffeomorphism $\varphi\colon \mathbb{R}^n \to \mathbb{R}^n$, and x_0 a hyperbolic equilibrium point of (17.3), then in a neighborhood of x_0 (17.3) is topologically conjugate to the linearization $y_{t+1} = D\varphi(x_0)y_t$.

17.1.4 Structural Stability (Robustness)

17.1.4.1 Structurally Stable Differential Equations

1. Definition

The differential equation (17.1), i.e., the vector field $f\colon M \to \mathbb{R}^n$, is called *structurally stable* or *robust*, if small perturbations of f result in topologically equivalent differential equations. The precise definition of robustness requires the notion of distance between two vector fields defined on M. The further investigations are restricted to smooth vector fields on M, which have a common open connected absorbing set $U \subset M$. Let the boundary ∂U of U be a smooth $(n-1)$-dimensional hypersurface and suppose that it can be represented as $\partial U = \{x \in \mathbb{R}^n\colon h(x) = 0\}$, where $h\colon \mathbb{R}^n \to \mathbb{R}$ is a C^1-function with $\operatorname{grad} h(x) \neq 0$ in a neighborhood of ∂U. Let $\mathrm{X}^1(U)$ be the metric space of all smooth vector fields on M with the C^1 *metric*

$$\rho(f,g) = \sup_{x\in U} \| f(x) - g(x)\| + \sup_{x\in U} \| D f(x) - D g(x)\|. \tag{17.25}$$

(In the first term of the right-hand side $\|\cdot\|$ means the Euclidean vector norm, in the second one the operator norm.) The smooth vector fields f intersecting transversally the boundary ∂U in the direction U, i.e., for which $\operatorname{grad} h(x)^T f(x) \neq 0$, $(x \in \partial U)$ and $\varphi^t(x) \in U$ $(x \in \partial U,\ t > 0)$ hold, form the set $\mathrm{X}^1_+(U) \subset \mathrm{X}^1(U)$. The vector field $f \in \mathrm{X}^1_+(U)$ is called *structurally stable* if there is a $\delta > 0$ such that every other vector field $g \in \mathrm{X}^1_+(U)$ with $\rho(f,g) < \delta$ is topologically equivalent to f.

■ Consider the planar differential equation $g(\cdot,\alpha)$

$$\dot{x} = -y + x(\alpha - x^2 - y^2), \quad \dot{y} = x + y(\alpha - x^2 - y^2) \tag{17.26}$$

with parameter α, where $|\alpha| < 1$. The differential equation g belongs, e.g., to $X^1_+(U)$ with $U = \{(x, y): x^2 + y^2 < 2\}$ **(Fig. 17.12a)**. Obviously, $\rho(g(\cdot, 0), g(\cdot, \alpha)) = |\alpha|(\sqrt{2} + 1)$. The vector field $g(\cdot, 0)$ is structurally unstable; there exist vector fields arbitrarily close to $g(\cdot, 0)$, which are not topologically equivalent to $g(\cdot, 0)$ **(Fig. 17.12b,c)**. This is clear, if the polar coordinate representation $\dot{r} = -r^3 + \alpha r$, $\dot{\vartheta} = 1$ of (17.26) is considered. For $\alpha > 0$ there always exists a stable limit cycle $r = \sqrt{\alpha}$.

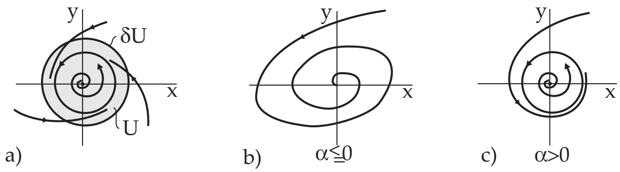


Figure 17.12

2. Structurally Stable Systems in the Plane

Suppose the planar differential equation (17.1) with $f \in X^1_+(U)$ is structurally stable. Then:

a) (17.1) has only a finite number of equilibrium points and periodic orbits.

b) All ω-limit sets $\omega(x)$ with $x \in \overline{U}$ of (17.1) consist of equilibrium points and periodic orbits only.

Theorem of Andronov and Pontryagin: The planar differential equation (17.1) with $f \in X^1_+(U)$ is structurally stable if and only if:

a) All equilibrium points and periodic orbits in $\overline{U}$ are hyperbolic.

b) There are no separatrices, i.e., no heteroclinic or homoclinic orbits, coming from a saddle and tending to a saddle point.

17.1.4.2 Structurally Stable Time Discrete Systems

In the case of time discrete systems (17.3), i.e., of mappings $\varphi: M \to M$, let $U \subset M \subset \mathbb{R}^n$ be a bounded, open, and connected set with a smooth boundary. Let $\mathrm{Diff}^1(U)$ be the metric space of all diffeomorphisms on M with the corresponding U defined C^1-metric. Suppose the set $\mathrm{Diff}^1_+(U) \subset \mathrm{Diff}(U)$ consists of the diffeomorphisms φ, for which $\varphi(\overline{U}) \subset U$ is valid. The mapping $\varphi \in \mathrm{Diff}^1_+(U)$ (and the corresponding dynamical system (17.3)) is called *structurally stable* if there exists a $\delta > 0$ such that every other mapping $\psi \in \mathrm{Diff}^1_+(U)$ with $\rho(\varphi, \psi) < \delta$ is topologically conjugate to φ.

17.1.4.3 Generic Properties

1. Definition

A property of elements of a metric space (M, ρ) is called *generic* (or *typical*) if the set of the elements B of M with this property form a *set of the second Baire category*, i.e., it can be represented as $B = \bigcap_{m=1,2\ldots} B_m$, where every set B_m is open and dense in M.

■ **A:** The sets $\mathbb{R}$ and $\mathbb{I} \subset \mathbb{R}$ (irrational numbers) are sets of second Baire category, but $\mathbb{Q} \subset \mathbb{R}$ is not.

■ **B:** Density alone as a property of "typical" is not enough: $\mathbb{Q} \subset \mathbb{R}$ and $\mathbb{I} \subset \mathbb{R}$ are both dense, but they cannot be typical at the same time.

■ **C:** There is no connection between the Lebesgue measure λ (see 12.9.1, **2.**, p. 694) of a set from $\mathbb{R}$ and the Baire category of this set. The set $B = \bigcap_{k=1,2,\ldots} B_k$ with $B_k = \bigcup_{n\geq 0} \left(a_n - \frac{1}{k\,2^n}, a_n + \frac{1}{k\,2^n}\right)$, where $\mathbb{Q} = \{a_n\}_{n=0}^{\infty}$ represents the rational numbers, is a set of second Baire category (see [17.5], [17.10]).

On the other hand, since $B_k \supset B_{k+1}$ and $\lambda(B_k) < +\infty$ also $\lambda(B) = \lim\limits_{k\to\infty} \lambda(B_k) \leq \lim\limits_{k\to\infty} \frac{2}{k}\frac{1}{1-1/2} = 0$ holds.

2. Generic Properties of Planar Systems, Hamiltonian Systems

For planar differential equations the set of all structurally stable systems from $\mathrm{X}^1_+(U)$ is open and dense in $\mathrm{X}^1_+(U)$. Hence, structurally stable systems are typical for the plane. It is also typical that every orbit of a planar system from $X^1_+(U)$ for increasing time tends to one of a finite number of equilibrium points and periodic orbits. Quasiperiodic orbits are not typical. Under certain assumptions, in the case of Hamiltonian systems, the quasiperiodic orbits of differential equation are preserved in the case of small perturbations. Hence, Hamiltonian systems are not typical systems.

■ Given in $\mathbb{R}^4$ a Hamiltonian system in action-angle variables $\dot{j}_1 = 0,\ \dot{j}_2 = 0,\ \dot{\Theta}_1 = \frac{\partial H_0}{\partial j_1},\ \dot{\Theta}_2 = \frac{\partial H_0}{\partial j_2}$, where the Hamiltonian $H_0(j_1, j_2)$ is analytical. Obviously, this system has the solutions $j_1 = c_1, j_2 = c_2, \Theta_1 = \omega_1 t + c_3,\ \Theta_2 = \omega_2 t + c_4$ with constants $c_1, \ldots, c_4$, where ω_1 and ω_2 can depend on c_1 and c_2. The relation $(j_1, j_2) = (c_1, c_2)$ defines an invariant torus T^2. Consider now the perturbed Hamiltonian

$$H_0(j_1, j_2) + \varepsilon H_1(j_1, j_2, \Theta_1, \Theta_2)$$

instead of H_0, where H_1 is analytical and $\varepsilon > 0$ is a small parameter.

The *Kolmogorov-Arnold-Moser theorem* (*KAM theorem*) says in this case that if H_0 is non-degenerate, i.e., $\det\left(\frac{\partial^2 H_0}{\partial j_k^2}\right) \neq 0$, then in the perturbed Hamiltonian system most of the invariant non-resonant tori will not vanish for sufficiently small $\varepsilon > 0$ but will be only slightly deformed. "Most of the tori" means that the Lebesgue measure of the complement set with respect to the tori tends to zero if ε tends to 0. A torus, defined as above and characterized by ω_1 and ω_2, is called non-resonant if there exists a constant $c > 0$ such that the inequality $\left|\frac{\omega_1}{\omega_2} - \frac{p}{q}\right| \geq \frac{c}{q^{2.5}}$ holds for all positive integers p and q.

3. Non-Wandering Points, Morse-Smale Systems

Let $\{\varphi^t\}_{t\in\mathrm{R}}$ be a dynamical system on the n-dimensional compact orientable manifold M. The point $p \in M$ is called *non-wandering* with respect to $\{\varphi^t\}$ if

$$\forall\, T > 0 \quad \exists\, t, \quad |t| \geq T: \quad \varphi^t(U_p) \cap U_p \neq \emptyset \tag{17.27}$$

holds for an arbitrary neighborhood $U_p \subset M$ of p.

■ Steady states and periodic orbits consist only of non-wandering points.

The set $\Omega(\varphi^t)$ of all non-wandering points of the dynamical systems generated by (17.1) is closed, invariant under $\{\varphi^t\}$ and contains all periodic orbits and all ω-limit sets of points from M.

The dynamical system $\{\varphi^t\}_{t\in\mathrm{R}}$ on M generated by a smooth vector field is called a Morse-Smale system if the following conditions are fulfilled:

1. The system has finitely many equilibrium points and periodic orbits and they are all hyperbolic.

2. All stable and unstable manifolds of equilibrium points and periodic orbits are transversal to each other.

3. The set of all non-wandering points consists only of equilibrium points and periodic orbits.

Theorem of Palis and Smale: Morse-Smale systems are structurally stable.

The converse statement of the theorem of Palis and Smale is not true: In the case of $n \geq 3$, there exist structurally stable systems with infinitely many periodic orbits.

For $n \geq 3$, structurally stable systems are not typical.

17.2 Quantitative Description of Attractors

17.2.1 Probability Measures on Attractors

17.2.1.1 Invariant Measure

1. Definition, Measure Concentrated on the Attractor

Let $\{\varphi^t\}_{t\in\Gamma}$ be a dynamical system on (M,ρ). Let $\mathcal{B}$ be the σ-algebra of Borel sets on M (12.9.1, **2.**, p. 694) and let $\mu\colon \mathcal{B} \to [0,+\infty]$ be a measure on $\mathcal{B}$. Every mapping φ^t is supposed to be μ measurable. The measure μ is called *invariant* under $\{\varphi^t\}_{t\in\Gamma}$ if $\mu(\varphi^{-t}(A)) = \mu(A)$ holds for all $A \in \mathcal{B}$ and $t > 0$. If the dynamical system $\{\varphi^t\}_{t\in\Gamma}$ is invertible, then the property of the measure being invariant under the dynamical system can be expressed as $\mu(\varphi^t(A)) = \mu(A) \quad (A \in \mathcal{B},\ t > 0)$. The measure μ is said to be *concentrated on the Borel set* $A \subset M$ if $\mu(M \setminus A) = 0$. If Λ is also an attractor of $\{\varphi^t\}_{t\in\Gamma}$ and μ is an invariant measure under $\{\varphi^t\}$, then it is concentrated at Λ, if $\mu(B) = 0$ for every Borel set B with $\Lambda \cap B = \emptyset$.

The *support* of a measure $\mu\colon \mathcal{B} \to [0,+\infty]$, denoted by supp μ, is the smallest closed subset of M on which the measure μ is concentrated.

■ **A:** The *Bernoulli shift mapping* is considered on $M = [0,1]$:

$$x_{t+1} = 2x_t \pmod 1. \tag{17.28a}$$

In this case the map $\varphi\colon [0,1] \to [0,1]$ is defined as

$$\varphi(x) = \begin{cases} 2x, & 0 \le x \le 1/2, \\ 2x-1, & 1/2 < x \le 1. \end{cases} \tag{17.28b}$$

The definition yields that the Lebesgue measure is invariant under the Bernoulli shift mapping. If a number $x \in [0,1)$ is written in dyadic form $x = \sum\limits_{n=1}^{\infty} a_n \cdot 2^{-n} \quad (a_n = 0 \text{ or } 1)$, then this representation can be identified with $x = .a_1a_2a_3\ldots$. The result of the operation $2x(\bmod 1)$ can be written as $.a_1'\,a_2'\,a_3'\ldots$ with $a_i' = a_{i+1}$, i.e., all digits a_k are shifted to the left by one position and the first digit is omitted.

■ **B:** The mapping $\Psi\colon [0,1] \to [0,1]$ with

$$\Psi(y) = \begin{cases} 2y, & 0 \le y < 1/2, \\ 2(1-y), & 1/2 \le y \le 1 \end{cases} \tag{17.29}$$

is called a *tent mapping* and the Lebesgue measure is an invariant measure. The homeomorphism $h\colon [0,1) \to [0,1)$ with $y = \dfrac{2}{\pi}\arcsin\sqrt{x}$ transforms the mapping φ from (17.5) into (17.29). Hence, in the case of $\alpha = 4$, (17.5) has an invariant measure which is absolutely continuous. For the density $\rho_1(y) \equiv 1$ of (17.29) and $\rho(x)$ of (17.5) at $\alpha = 4$ it is valid that $\rho_1(y) = \rho(h^{-1}(y))\,|(h^{-1})'(y)|$. It follows directly that $\rho(x) = \dfrac{1}{\pi\sqrt{x(1-x)}}$.

■ **C:** If x_0 is a stable periodic point of period T of the invertible discrete dynamical system $\{\varphi^i\}$, then $\mu = \dfrac{1}{T}\sum\limits_{i=0}^{T-1} \delta_{\varphi^i(x_0)}$ is a probability measure for $\{\varphi^i\}$. Here, δ_{x_0} is the *Dirac measure* concentrated at x_0 (see 12.9.1,**2.**, p. 694).

2. Natural Measure

Let Λ be an attractor of $\{\varphi^t\}_{t\in\Gamma}$ in M with domain of attraction W. For an arbitrary Borel set $A \subset W$ and an arbitrary point $x_0 \in W$ define the number:

$$\mu(A;x_0) := \lim_{T\to\infty} \frac{t(T,A,x_0)}{T}. \tag{17.30}$$

Here, $t(T, A, x_0)$ is that part of the time $T > 0$ for which the orbit portion $\{\varphi^t(x_0)\}_{t=0}^{T}$ lies in the set A. If $\mu(A; x_0) = \alpha$ for λ-a.e.* x_0 from W, then let $\mu(A) := \mu(A; x_0)$. Since almost all orbits with initial points $x_0 \in W$ tend to Λ for $t \to +\infty$, μ is a probability measure concentrated at Λ.

17.2.1.2 Elements of Ergodic Theory

1. Ergodic Dynamical Systems

A dynamical system $\{\varphi^t\}_{t\in\Gamma}$ on (M, ρ) with invariant measure μ is called *ergodic* (one also says that the measure is ergodic) if either $\mu(A) = 0$ or $\mu(M \setminus A) = 0$ for every Borel set A with $\varphi^{-t}(A) = A\,(\forall\, t > 0)$. If $\{\varphi^t\}$ is a discrete dynamical system (17.3), $\varphi\colon M \to M$ is a homeomorphism and M is a compact metric space, then there always exists an invariant ergodic measure.

■ **A:** Suppose there is given the *rotation mapping of the circle* S^1

$$x_{t+1} = x_t + \Phi \pmod{2\pi}, \quad t = 0, 1, \dots , \tag{17.31}$$

with $\varphi\colon [0, 2\pi) \to [0, 2\pi)$, defined by $\varphi(x) = x + \Phi\,(\bmod\, 2\pi)$. The Lebesgue measure is invariant under φ. If $\dfrac{\Phi}{2\pi}$ is irrational, then (17.31) is ergodic; if $\dfrac{\Phi}{2\pi}$ is rational, then (17.31) is not ergodic.

■ **B:** Dynamical systems with stable equilibrium points or stable periodic orbits as attractors are ergodic with respect to the natural measure.

Birkhoff Ergodic Theorem: Suppose that the dynamical system $\{\varphi^t\}_{t\in\Gamma}$ is ergodic with respect to the invariant probability measure μ. Then, for every integrable function $h \in L^1(M, \mathcal{B}, \mu)$, the time average along the positive semiorbits $\{\varphi^t(x_0)\}_{t=0}^{\infty}$, i.e. $\overline{h}(x_0) = \lim\limits_{T\to+\infty} \dfrac{1}{T}\displaystyle\int_0^T h\left(\varphi^t(x_0)\right) dt$ for flows and $\overline{h}(x_0) = \lim\limits_{n\to\infty} \dfrac{1}{n}\displaystyle\sum_{i=0}^{n-1} h\left(\varphi^i(x_0)\right)$ for discrete systems, coincide with the space average $\displaystyle\int_M h\, d\mu$ for μ-a.e. points $x_0 \in M$.

2. Physical or SBR Measure

The statement of the ergodic theorem is useful only if the support of the measure μ is large. Let $\varphi\colon M \to M$ be a continuous mapping, and $\mu\colon \mathcal{B} \to \mathbb{R}$ be an invariant measure. The measure μ is called a *SBR measure* (according to Sinai, Bowen and Ruelle, see also [17.9]) if for any continuous function $h\colon M \to \mathbb{R}$ the set of all points $x_0 \in M$, for which

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} h(\varphi^i(x_0)) = \int_M h\, d\mu \tag{17.32a}$$

holds, has a positive Lebesgue measure for this. It is sufficient that the sequence of measures

$$\mu_n := \frac{1}{n} \sum_{i=0}^{n-1} \delta_{\varphi^i(x)} \tag{17.32b}$$

where δ_x is the Dirac measure, weakly converges to μ for almost all $x \in M$, i.e., for every continuous function $\displaystyle\int_M h\, d\mu_n \to \int_M h d\mu$ as $n \to +\infty$.

■ For some important attractors, such as the Hénon attractor, the existence of an SBR measure is proven.

3. Mixing Dynamical Systems

A dynamical system $\{\varphi^t\}_{t\in\Gamma}$ on (M, ρ) with invariant probability measure μ is called *mixing* if $\lim\limits_{t\to+\infty} \mu\left(A \cap \varphi^{-t}(B)\right) = \mu(A)\mu(B)$ holds for arbitrary Borel sets $A, B \subset M$. For a mixing system, the measure of the set of all points which are at $t = 0$ in A and under φ^t for large t in B, depends only on

*Here and in the following a.e. is an abbreviation for "almost everywhere".

the product $\mu(A)\mu(B)$.
A mixing system is also ergodic: Let $\{\varphi^t\}$ be a mixing system and A be a Borel set with $\varphi^{-t}(A) = A$ $(t > 0)$. Then $\mu(A)^2 = \lim_{t\to\infty} \mu(\varphi^{-t}(A) \cap A) = \mu(A)$ holds and $\mu(A)$ is 0 or 1.
A flow $\{\varphi^t\}$ of (17.1) is mixing if and only if the relation

$$\lim_{t\to+\infty} \int_M [g(\varphi^t(x)) - \overline{g}\,][h(x) - \overline{h}\,]\, d\mu = 0 \tag{17.33}$$

holds for arbitrary quadratically integrable functions $g, h \in L^2(M, \mathcal{B}, \mu)$. Here, $\overline{g}$ and $\overline{h}$ denote the space average, which is replaced by the time average.

■ The modulo mapping (17.28a) is mixing. The rotation mapping (17.31) is not mixing with respect to the probability measure $\dfrac{\lambda}{2\pi}$.

4. Autocorrelation Function

Suppose the dynamical system $\{\varphi^t\}_{t\in\Gamma}$ on M with invariant measure μ is ergodic. Let $h\colon M \to \mathbb{R}$ be an arbitrary continuous function, $\{\varphi^t(x)\}_{t\geq 0}$ be an arbitrary semiorbit and let the space average $\overline{h}$ be replaced by the time average, i.e., by $\lim_{T\to\infty} \dfrac{1}{T}\displaystyle\int_0^T h\left(\varphi^t(x)\right) dt$ in the time-continuous case and by $\lim_{n\to\infty} \dfrac{1}{n}\displaystyle\sum_{i=0}^{n-1} h\left(\varphi^i(x)\right)$ in the time-discrete case. With respect to h the *autocorrelation function* along the semiorbit $\{\varphi^t(x)\}_{t\geq 0}$ to a time point $\tau \geq 0$ is defined for a flow by

$$C_h(\tau) = \lim_{T\to\infty} \frac{1}{T} \int_0^T h\left(\varphi^{t+\tau}(x)\right) h\left(\varphi^t(x)\right) dt - \overline{h}^{\,2} \tag{17.34a}$$

and for a discrete system by

$$C_h(\tau) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} h\left(\varphi^{i+\tau}(x)\right) h\left(\varphi^i(x)\right) - \overline{h}^{\,2}. \tag{17.34b}$$

The autocorrelation function is defined also for negative time, where $C_h(\cdot)$ is considered as an even function on $\mathbb{R}$ or $\mathbb{Z}$.

Periodic or quasiperiodic orbits lead to periodic or quasiperiodic behavior of C_h. A quicker descent of $C_h(\tau)$ for increasing τ and arbitrary test function h refers to chaotic behavior. If $C_h(\tau)$ decreases for increasing τ with an exponential speed, then it means mixed behavior.

5. Power Spectrum

The Fourier transform of $C_h(\tau)$ is called a *power spectrum* (see also 15.3.1.2, **5.**, p. 786) and is denoted by $P_h(\omega)$. In the time-continuous case, under the assumption that $\int_{-\infty}^{+\infty} |C_h(\tau)| d\tau < \infty$,

$$P_h(\omega) = \int_{-\infty}^{+\infty} C_h(\tau) e^{-\mathrm{i}\omega\tau}\, d\tau = 2\int_0^{\infty} C_h(\tau)\cos(\omega\tau)\, d\tau. \tag{17.35a}$$

In the time-discrete case, if $\sum_{k=-\infty}^{+\infty} |C_h(k)| < +\infty$ holds, then

$$P_h(\omega) = C_h(0) + 2\sum_{k=1}^{\infty} C_h(k)\cos\omega k \tag{17.35b}$$

holds. If the absolute integrability or summability of $C_h(\cdot)$ does not hold, then, in the most important cases, P_h can be considered as a distribution. The power spectrum corresponding to the periodic

motions of a dynamical system is characterized by equidistant impulses. For quasiperiodic motions, there occur impulses in the power spectrum, which are linear combinations with integer coefficients of the basic impulses of the quasiperiodic motion. A "wide-band spectrum with singular peaks" can be considered as an indicator of chaotic behavior.

■ **A:** Let φ be a T-periodic orbit of (17.1), h be a test function such that the time average of $h(\varphi(t))$ is zero, and suppose $h(\varphi(t))$ has the Fourier representation

$$h(\varphi(t)) = \sum_{k=-\infty}^{+\infty} \alpha_k e^{\mathrm{i}k\omega_0 t} \quad \text{with} \quad \omega_0 = \frac{2\pi}{T}.$$

Then with δ as the δ distribution, holds

$$C_h(\tau) = \sum_{k=-\infty}^{+\infty} |\alpha_k|^2 \cos(k\omega_0\tau) \quad \text{and} \quad P_h(\omega) = 2\pi \sum_{k=-\infty}^{+\infty} |\alpha_k|^2 \delta(\omega - k\omega_0).$$

■ **B:** Suppose φ is a quasiperiodic orbit of (17.1), h is a test function such that the time average is zero along φ, and let $h(\varphi(t))$ be the representation (double Fourier series)

$$h(\varphi(t)) = \sum_{k_1=-\infty}^{+\infty} \sum_{k_2=-\infty}^{+\infty} \alpha_{k_1k_2} e^{\mathrm{i}(k_1\omega_1+k_2\omega_2)t}.$$

Then,

$$C_h(\tau) = \sum_{k_1=-\infty}^{+\infty} \sum_{k_2=-\infty}^{+\infty} |\alpha_{k_1k_2}|^2 \cos(k_1\omega_1 + k_2\omega_2)\tau,$$

$$P_h(\omega) = 2\pi \sum_{k_1=-\infty}^{+\infty} \sum_{k_2=-\infty}^{+\infty} |\alpha_{k_1k_2}|^2 \delta(\omega - k_1\omega_1 - k_2\omega_2).$$

17.2.2 Entropies

17.2.2.1 Topological Entropy

Let (M, ρ) be a compact metric space and $\{\varphi^k\}_{k\in\Gamma}$ be a continuous dynamical system with discrete time on M. A distance function ρ_n on M for arbitrary $n \in \mathbb{N}$ is defined by

$$\rho_n(x, y) := \max_{0\le i\le n} \rho\,(\varphi^i(x), \varphi^i(y)). \tag{17.36}$$

Furthermore, let $N(\varepsilon, \rho_n)$ be the largest number of points from M which have in the metric ρ_n a distance at least ε from each other. The *topological entropy* of the discrete dynamical system (17.3) or of the mapping φ is $h(\varphi) = \lim_{\varepsilon\to 0} \limsup_{n\to\infty} \frac{1}{n} \ln\, N(\varepsilon, \rho_n)$. The topological entropy is a measure for the complexity of the mapping. Let (M_1, ρ_1) be a further compact metric space and $\varphi_1\colon M_1 \to M_1$ be a continuous mapping. If both mappings φ and φ_1 are topologically conjugate, then their topological entropies coincide. In particular, the topological entropy does not depend on the metric. For arbitrary $n \in \mathbb{N}$, $h\,(\varphi^n) = n\,h\,(\varphi)$ holds. If φ is a homeomorphism, then $h\,(\varphi^k) = |k|\,h\,(\varphi)$ for all $k \in \mathbb{Z}$. Based on the last property the topological entropy $h(\varphi^t) := h\,(\varphi^1)$ is defined for a flow $\varphi^t = \varphi(t, \cdot)$ of (17.1) on $M \subset \mathbb{R}^n$.

17.2.2.2 Metric Entropy

Let $\{\varphi^t\}_{t\in\Gamma}$ be a dynamical system on M with attractor Λ and with an invariant probability measure μ concentrated on Λ. For an arbitrary $\varepsilon > 0$ consider the cubes $Q_1(\varepsilon), \dots, Q_{n(\varepsilon)}(\varepsilon)$ of the form $\{(x_1, \dots, x_n)\colon k_i\varepsilon \le x_i < (k_i + 1)\varepsilon\ \ (i = 1, 2, \dots, n)\}$ with $k_i \in \mathbb{Z}$, for which $\mu(Q_i) > 0$. For arbitrary x from a Q_i the semiorbit $\{\varphi^t(x)\}_{t=0}^{\infty}$ is followed for increasing t. In time-distances of $\tau > 0$ ($\tau = 1$ in discrete systems), the N cubes, in which the semiorbit is found is denoted by $i_1, \dots, i_N$ after each other. Let $E_{i_1,\dots,i_N}$ be the set of all starting points in the neighborhood of Λ whose semiorbits at the

times $t_i = i\tau$ $(i = 1, 2, \ldots, N)$ are always in $Q_{i_1}, \ldots, Q_{i_N}$ and let $p(i_1, \cdots, i_N) = \mu(E_{i_1,\cdots,i_N})$ be the probability that a (typical) starting point is in $E_{i_1,\cdots,i_N}$.

The entropy gives the increment of the information on average by an experiment which shows that among a finite number of disjoint events which one has really happened. In the above situation this is

$$H_N = -\sum_{(i_1,\cdots,i_N)} p(i_1, \cdots, i_N) \ln p(i_1, \cdots, i_N), \tag{17.37}$$

where the summation is over all symbol sequences $(i_1, \cdots, i_N)$ with length N, which are realized by the orbits described above.

The *metric entropy* or *Kolmogorov-Sinai entropy* h_μ of the attractor Λ of $\{\varphi^t\}$ with respect to the invariant measure μ is the quantity $h_\mu = \lim\limits_{\varepsilon\to 0} \lim\limits_{N\to\infty} \dfrac{H_N}{\tau N}$. For discrete systems, the limit as $\varepsilon \to 0$ is omitted. For the topological entropy $h(\varphi)$ of $\varphi\colon \Lambda \to \Lambda$ the inequality $h_\mu \le h(\varphi)$ holds. In several cases $h(\varphi) = \sup\{h_\mu\colon \mu$ invariant probability measure on $\Lambda\}$.

■ **A:** Suppose $\Lambda = \{x_0\}$ is a stable equilibrium point of (17.1) as an attractor, with the natural measure μ concentrated on x_0. For these attractors $h_\mu = 0$.

■ **B:** For the shift mapping (17.28a), $h(\varphi) = h_\mu = \ln 2$, where μ is the invariant Lebesgue measure.

17.2.3 Lyapunov Exponents

1. Singular Values of a Matrix

Let L be an arbitrary matrix of type (n, n). The *singular values* $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_n$ of L are the non-negative roots of the eigenvalues $\alpha_1 \ge \cdots \ge \alpha_n \ge 0$ of the positive semidefinite matrix $L^{\mathrm{T}}L$. The eigenvalues α_i are enumerated according to their multiplicity.

The singular values can be interpreted geometrically. If K_ε is a sphere with center at 0 and with radius $\varepsilon > 0$, then the image $L(K_\varepsilon)$ is an ellipsoid with semi-axis lengths $\sigma_i\varepsilon$ $(i = 1, 2, \ldots, n)$ **(Fig. 17.13a)**.

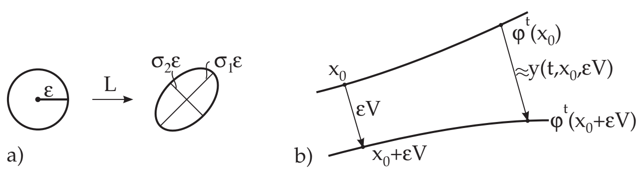


Figure 17.13

2. Definition of Lyapunov Exponents

Let $\{\varphi^t\}_{t\in\Gamma}$ be a smooth dynamical system on $M \subset \mathbb{R}^n$, which has an attractor Λ with an invariant ergodic probability measure μ concentrated on Λ. Let $\sigma_1(t, x) \ge \cdots \ge \sigma_n(t, x)$ be the singular values of the Jacobian matrix $D\,\varphi^t(x)$ of φ^t at the point x for arbitrary $t \ge 0$ and $x \in \Lambda$. Then there exists a sequence of numbers $\lambda_1 \ge \cdots \ge \lambda_n$, the *Lyapunov exponents*, such that $\dfrac{1}{t} \ln \sigma_i(t, x) \to \lambda_i$ for $t \to +\infty$ μ-a.e. in the sense of L^1. According to the theorem of Oseledec, there exists μ-a.e. a sequence of subspaces of $\mathbb{R}^n$

$$\mathbb{R}^n = E^x_{s_1} \supset E^x_{s_2} \supset \cdots \supset E^x_{s_{r+1}} = \{0\}, \tag{17.38}$$

such that for μ-a.e. x the quantity $\dfrac{1}{t} \ln \|D\,\varphi^t(x)v\|$ tends to an element $\lambda_{s_j} \in \{\lambda_1, \ldots, \lambda_n\}$ uniformly with respect to $v \in E^x_{s_j} \setminus E^x_{s_{j+1}}$.

3. Calculation of Lyapunov Exponents

Suppose $\sigma_i(t,x)$ are the semi-axis lengths of the ellipsoid got by deformation of the unit sphere with center at x by $D\,\varphi^t(x)$. The formula $\chi_i(x) = \limsup\limits_{t\to\infty} \frac{1}{t} \ln \sigma_i(t,x)$ can be used to calculate the Lyapunov exponents, if additionally a reorthonormalization method, such as Householder, is used. The function $y(t,x,v) = D\,\varphi^t(x)v$ is the solution of the variational equation with v at $t=0$ associated to the semiorbit $\gamma^+(x)$ of the flow $\{\varphi^t\}$. Actually, $\{\varphi^t\}_{t\in\mathbb{R}}$ is the flow of (17.1), so the variational equation is $\dot{y} = D\,f(\varphi^t(x))\,y$. The solution of this equation with initial v at time $t=0$ can be represented as $y(t,x,v) = \Phi_x(t)v$, where $\Phi_x(t)$ is the normed fundamental matrix of the variational equation at $t=0$, which is a solution of the matrix differential equation $\dot{Z} = D\,f(\varphi^t(x))\,Z$ with initial $Z(0)=E_n$ according to the theorem about differentiability with respect to the initial state (see 17.1.1.1, **2.**, p. 857).

The number $\chi(x,v) = \limsup\limits_{t\to\infty} \frac{1}{t} \ln \|D\,\varphi^t(x)v\|$ describes the behavior of the orbit $\gamma(x+\varepsilon v), 0<\varepsilon\ll 1$ with initial $x+\varepsilon v$ with respect to the initial orbit $\gamma(x)$ in the direction v. If $\chi(x,v)<0$, then the orbits move nearer to x for increasing t in the direction v. If, on the contrary, $\chi(x,v)>0$, then the orbits move away **(Fig. 17.13b)**.

Let Λ be the attractor of the dynamical system $\{\varphi^t\}_{t\in\Gamma}$ and μ the invariant ergodic measure concentrated on it. Then, the sum of all Lyapunov exponents μ-a.e. $x\in\Lambda$ is

$$\sum_{i=1}^{n} \lambda_i = \lim_{t\to\infty} \frac{1}{t} \int_0^t \operatorname{div} f(\varphi^s(x))\,ds \tag{17.39a}$$

in the case of flows (17.1) and for a discrete system (17.3), it is

$$\sum_{i=1}^{n} \lambda_i = \lim_{k\to\infty} \frac{1}{k} \sum_{i=0}^{k-1} \ln|\det D\,\varphi(\varphi^i(x))|. \tag{17.39b}$$

Hence, in dissipative systems $\sum\limits_{i=1}^{n} \lambda_i < 0$ holds. Considering that one of the Lyapunov exponents is equal to zero if the attractor is not an equilibrium point, the calculation of Lyapunov exponents can be simplified (see [17.9]).

■ **A:** Let be x_0 an equilibrium point of the flow of (17.1) and let α_i be the eigenvalues of the Jacobian matrix at x_0. With the measure concentrated on x_0, the following holds for the Lyapunov exponents: $\lambda_i = \operatorname{Re}\alpha_i \ \ (i=1,2,\ldots,n)$.

■ **B:** Let $\gamma(x_0) = \{\varphi^t(x_0),\ t\in[0,T]\}$ be a T-periodic orbit of (17.1) and let ρ_i be the multipliers of $\gamma(x_0)$. With the measure concentrated on $\gamma(x_0)$ there is $\lambda_i = \frac{1}{T}\ln|\rho_i|$ for $i=1,2\ldots,n$.

4. Metric Entropy and Lyapunov Exponents

If $\{\varphi^t\}_{t\in\Gamma}$ is a dynamical system on $M\subset\mathbb{R}^n$ with attractor Λ and an ergodic probability measure μ concentrated on Λ, then the inequality $h_\mu \le \sum\limits_{\lambda_i>0} \lambda_i$ holds for the metric entropy h_μ, where in the sum the Lyapunov exponents are repeated according to their multiplicity.

The equality

$$h_\mu = \sum_{\lambda_i>0} \lambda_i \quad \text{(Pesin's formula)} \tag{17.40}$$

is not valid in general (see also 17.2.4.4, ■ **B**, p. 888). If the measure μ is absolutely continuous with respect to the Lebesgue measure and $\varphi\colon M\to M$ is a C^2-diffeomorphism, then Pesin's formula is valid.

17.2.4 Dimensions

17.2.4.1 Metric Dimensions

1. Fractals

Attractors or other invariant sets of dynamical systems can be geometrically more complicated than point, line or torus. *Fractals*, independently of dynamics, are sets which distinguish themselves by one or several characteristics such as fraying, porosity, complexity, and self-similarity. Since the usual notion of dimension used for smooth surfaces and curves cannot be applied to fractals, a generalized definition of the dimension is necessary. For more details see [17.2], [17.12].

■ The interval $G_0 = [0,1]$ is divided into three subintervals with the same length and the middle open third is removed, so the set $G_1 = [0, \frac{1}{3}] \cup [\frac{2}{3}, 1]$ is obtained. Then from both subintervals of G_1 the open middle third ones are removed, which yields the set $G_2 = \left[0, \frac{1}{9}\right] \cup \left[\frac{2}{9}, \frac{1}{3}\right] \cup \left[\frac{2}{3}, \frac{7}{9}\right] \cup \left[\frac{8}{9}, 1\right]$. Continuing this procedure, G_k is obtained from G_{k-1} by removing the open middle thirds from the subintervals. So, a sequence of sets $G_0 \supset G_1 \supset \cdots \supset G_n \supset \cdots$ is obtained, where every G_n consists of 2^n intervals of length $\dfrac{1}{3^n}$.

The *Cantor set* C is defined as the set of all points belonging to all G_n, i.e., $C = \bigcap\limits_{n=1}^{\infty} G_n$. The set C is compact, uncountable, its Lebesgue measure is zero and it is perfect, i.e., C is closed and every point is an accumulation point. The Cantor set can be an example of a fractal.

2. Hausdorff Dimension

The motivation for this dimension comes from volume calculation based on Lebesgue measure. If supposing that a bounded set $A \subset \mathbb{R}^3$ is covered by a finite number of spheres B_{r_i} with radius $r_i \le \varepsilon$, so that $\bigcup_i B_{r_i} \supset A$ holds, then for A a "rough volume" is $\sum\limits_i \dfrac{4}{3}\pi r_i^3$. Now, one defines the quantity $\mu_\varepsilon(A) = \inf\{\sum\limits_i \dfrac{4}{3}\pi r_i^3\}$ over all finite coverings of A by spheres with radius $r_i \le \varepsilon$. If ε tends to zero, then one gets the Lebesgue outer measure $\overline{\lambda}(A)$ of A. If A is measurable, the outer measure is equal to the volume vol(A).

Suppose M is the Euclidean space $\mathbb{R}^n$ or, more generally, a separable metric space with metric ρ and let $A \subset M$ be a subset of it. For arbitrary parameters $d \ge 0$ and $\varepsilon > 0$, the quantity

$$\mu_{d,\varepsilon}(A) = \inf\left\{\sum_i (\mathrm{diam} B_i)^d : A \subset \bigcup B_i,\ \mathrm{diam} B_i \le \varepsilon\right\} \tag{17.41a}$$

is determined, where $B_i \subset M$ are arbitrary subsets with diameter $\mathrm{diam} B_i = \sup\limits_{x,y \in B_i} \rho(x,y)$.

The *Hausdorff outer measure of dimension d of A* is defined by

$$\mu_d(A) = \lim_{\varepsilon \to 0} \mu_{d,\varepsilon}(A) = \sup_{\varepsilon > 0} \mu_{d,\varepsilon}(A) \tag{17.41b}$$

and it can be either finite or infinite. The *Hausdorff dimension* $d_H(A)$ of the set A is then the (unique) critical value of the Hausdorff measure:

$$d_H(A) = \begin{cases} +\infty, & \text{if } \mu_d(A) \ne 0 \text{ for all } d \ge 0, \\ \inf\{d \ge 0\colon \mu_d(A) = 0\}. \end{cases} \tag{17.41c}$$

Remark: The quantities $\mu_{d,\varepsilon}(A)$ can be defined with coverings of spheres with radius $r_i \le \varepsilon$ or, in the case of $\mathbb{R}^n$, of cubes with side length $\le \varepsilon$.

Important Properties of the Hausdorff Dimension:

(HD1) $d_H(\emptyset) = 0$.

(HD2) If $A \subset \mathbb{R}^n$, then $0 \le d_H(A) \le n$.

(HD3) From $A \subset B$, it follows that $d_H(A) \le d_H(B)$.

(HD4) If $A = \bigcup_{i=1}^{\infty} A_i$, then $d_H(A) = \sup_i d_H(A_i)$.

(HD5) If A is finite or countable, then $d_H(A) = 0$.

(HD6) If $\varphi\colon M \to M$ is Lipschitz continuous, i.e., there exists a constant $L > 0$ such that $\rho(\varphi(x), \varphi(y)) \le L\rho(x, y)\, \forall x, y \in M$, then $d_H(\varphi(A)) \le d_H(A)$. If the inverse mapping φ^{-1} exists as well, and it is also Lipschitz continuous, then $d_H(A) = d_H(\varphi(A))$.

■ For the set $\mathbb{Q}$ of all rational numbers $d_H(\mathbb{Q}) = 0$ because of **(HD5)**. The dimension of the Cantor set C is $d_H(C) = \dfrac{\ln 2}{\ln 3} \approx 0.6309\ldots$.

3. Box-Counting Dimension or Capacity

Let A be a compact set of the metric space (M, ρ) and let $N_\varepsilon(A)$ be the minimal number of sets of diameter $\le \varepsilon$, necessary to cover A. The quantity

$$\overline{d}_B(A) = \limsup_{\varepsilon \to 0} \frac{\ln N_\varepsilon(A)}{\ln \frac{1}{\varepsilon}} \tag{17.42a}$$

is called the *upper box-counting dimension* or *upper capacity*, the quantity

$$\underline{d}_B(A) = \liminf_{\varepsilon \to 0} \frac{\ln N_\varepsilon(A)}{\ln \frac{1}{\varepsilon}} \tag{17.42b}$$

is called the *lower box-counting dimension* or *lower capacity* (then d_C) of A. If $\overline{d}_B(A) = \underline{d}_B(A) := d_B(A)$ holds, then $d_B(A)$ is called the *box-counting dimension* of A. In $\mathbb{R}^n$ the box-counting dimension can be considered also for bounded sets which are not closed.

For a bounded set $A \subset \mathbb{R}^n$, the number $N_\varepsilon(A)$ in the above definition can also be defined in the following way: Let $\mathbb{R}^n$ be covered by a grid from n-dimensional cubes with side length ε. Then, $N_\varepsilon(A)$ can be the number of cubes of the grid having a non-empty intersecting A.

Important Properties of the Box-Counting Dimension:

(BD1) $d_H(A) \le d_B(A)$ always holds.

(BD2) For m-dimensional surfaces $F \subset \mathbb{R}^n$ holds $d_H(F) = d_B(F) = m$.

(BD3) $d_B(A) = d_B(\overline{A})$ holds for the closure $\overline{A}$ of A, while often $d_H(A) < d_H(\overline{A})$ is valid.

(BD4) If $A = \bigcup_n A_n$, then, in general, the formula $d_B(A) = \sup_n d_B(A_n)$ does not hold for the box-counting dimension.

■ Suppose $A = \{0, 1, \frac{1}{2}, \frac{1}{3}, \ldots\}$. Then $d_H(A) = 0$ and $d_B(A) = \dfrac{1}{2}$.

If A is the set of all rational points in $[0, 1]$, then because of **BD2** and **BD3** $d_B(A) = 1$ holds. On the other hand $d_H(A) = 0$.

4. Self-Similarity

Several geometric figures, which are called *self-similar*, can be derived by the following procedure: An initial figure is replaced by a new one which is composed of p copies of the original, any of them scaled linearly by a factor $q > 1$. All figures that are k times scalings of the initial figure in the k-th step are handled as the in the first step.



Figure 17.14

■ **A:** Cantor set: $p = 2$, $q = 3$.

■ **B:** Koch curve: $p = 4$, $q = 3$. The first three steps are shown in **Fig. 17.14**.

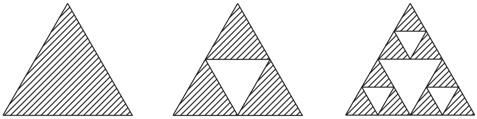


Figure 17.15

■ **C:** Sierpinski gasket: $p = 3$, $q = 2$. The first three steps are shown in **Fig. 17.15**. (The white triangles are always removed.)

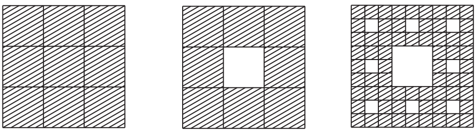


Figure 17.16

■ **D:** Sierpinski carpet: $p = 8$, $q = 3$. The first three steps are shown in **Fig. 17.16**. (The white squares are always removed.)

For the sets in the examples **A**–**D**:

$$d_B = d_H = \frac{\ln p}{\ln q}.$$

17.2.4.2 Dimensions Defined by Invariant Measures

1. Dimension of a Measure

Let μ be a probability measure in (M, ρ), concentrated on Λ. If $x \in \Lambda$ is an arbitrary point and $B_\delta(x)$ is a sphere with radius δ and center at x, then

$$\overline{d}_\mu(x) = \limsup_{\delta \to 0} \frac{\ln \mu(B_\delta(x))}{\ln \delta} \tag{17.43a}$$

denotes the *upper* and

$$\underline{d}_\mu(x) = \liminf_{\delta \to 0} \frac{\ln \mu(B_\delta(x))}{\ln \delta} \tag{17.43b}$$

denotes the *lower point-wise dimension* of μ in x. If $\underline{d}_\mu(x) = \overline{d}_\mu(x) := d_\mu(x)$, then $d_\mu(x)$ is called the *dimension of the measure* μ in x.

Young Theorem 1: If the relation $d_\mu(x) = \alpha$ holds for μ-a.e.*, $x \in \Lambda$, then

$$\alpha = d_H(\mu) := \inf_{X \subset \Lambda,\, \mu(X)=1} \{d_H(X)\}. \tag{17.44}$$

The quantity $d_H(\mu)$ is called the *Hausdorff dimension of the measure* μ.

■ Suppose $M = \mathbb{R}^n$ and let $\Lambda \subset \mathbb{R}^n$ be a compact sphere with Lebesgue measure $\lambda(\Lambda) > 0$. The restriction of μ to Λ is $\mu_\Lambda = \dfrac{\lambda}{\lambda(\Lambda)}$. Then

$\mu(B_\delta(x)) \sim \delta^n$ and $d_H(\mu) = n$.

2. Information Dimension

Suppose, the attractor Λ of $\{\varphi^t\}_{t \in \Gamma}$ is covered by cubes $Q_1(\varepsilon), \ldots, Q_{n(\varepsilon)}(\varepsilon)$ of side length ε as in 17.2.2.2, p. 879. Let μ be an invariant probability measure on Λ.
The *entropy* of the covering $Q_1(\varepsilon), \ldots, Q_{n(\varepsilon)}(\varepsilon)$ is

$$H(\varepsilon) = -\sum_{i=1}^{n(\varepsilon)} p_i(\varepsilon) \ln p_i(\varepsilon), \quad \text{with} \quad p_i(\varepsilon) = \mu(Q_i(\varepsilon)) \quad (i = 1, \ldots, n(\varepsilon)). \tag{17.45}$$

If the limit $d_I(\mu) = -\lim\limits_{\varepsilon \to 0} \dfrac{H(\varepsilon)}{\ln \varepsilon}$ exists, then this quantity has the property of a dimension and is called the *information dimension*.

*Here and in the following a.e. is an abbreviation for "almost everywhere".

Young Theorem 2: If the relation $d_\mu(x) = \alpha$ holds for μ-a.e. $x \in \Lambda$, then

$$\alpha = d_H(\mu) = d_I(\mu). \tag{17.46}$$

■ **A:** Let the measure μ be concentrated at the equilibrium point x_0 of $\{\varphi^t\}$. Since $H_\varepsilon(\mu) = -1 \ln 1 = 0$ always holds for $\varepsilon > 0$, so $d_I(\mu) = 0$.

■ **B:** Suppose the measure μ is concentrated on the limit cycle of $\{\varphi^t\}$. For $\varepsilon > 0$, $H_\varepsilon(\mu) = -\ln\varepsilon$ holds and so $d_I(\mu) = 1$.

3. Correlation Dimension

Let $\{y_i\}_{i=1}^\infty$ be a typical sequence of points of the attractor $\Lambda \subset \mathbb{R}^n$ of $\{\varphi^t\}_{t\in\Gamma}$, μ an invariant probability measure on Λ and let $m \in \mathbb{N}$ be arbitrary. For the vectors $x_i := (y_i, \dots, y_{i+m})$ let the distance be defined as $\operatorname{dist}(x_i, x_j) := \max\limits_{0\le s\le m} \|y_{i+s} - y_{j+s}\|$, where $\|\cdot\|$ is the Euclidean vector norm. If Θ denotes the Heaviside function $\Theta(x) = \begin{cases} 0, & x \le 0, \\ 1, & x > 0, \end{cases}$ then the expression

$$C^m(\varepsilon) = \limsup_{N\to+\infty} \frac{1}{N^2} \operatorname{card}\{(x_i, x_j)\colon \operatorname{dist}(x_i, x_j) < \varepsilon\}$$

$$= \limsup_{N\to\infty} \frac{1}{N^2} \sum_{i,j=1}^{N} \Theta\left(\varepsilon - \operatorname{dist}(x_i, x_j)\right) \tag{17.47a}$$

is called the *correlation integral*. The quantity

$$d_K = \lim_{\varepsilon\to 0} \frac{\ln C^m(\varepsilon)}{\ln\varepsilon} \tag{17.47b}$$

(if it exists) is the *correlation dimension*.

4. Generalized Dimension

Let the attractor Λ of $\{\varphi^t\}_{t\in\Gamma}$ on M with invariant probability measure μ be covered by cubes with side length ε as in 17.2.2.2, p. 879. For an arbitrary parameter $q \in \mathbb{R}$, $q \ne 1$, the sum

$$H_q(\varepsilon) = \frac{1}{1-q} \ln \sum_{i=1}^{n(\varepsilon)} p_i(\varepsilon)^q \quad \text{where} \quad p_i(\varepsilon) = \mu(Q_i(\varepsilon)) \tag{17.48a}$$

is called the *generalized entropy of q-th order* with respect to the covering $Q_1(\varepsilon), \dots, Q_{n(\varepsilon)}(\varepsilon)$.
The *Rényi dimension of q-th order* is

$$d_q = -\lim_{\varepsilon\to 0} \frac{H_q(\varepsilon)}{\ln\varepsilon}, \tag{17.48b}$$

if this limit exists.

Special Cases of the Rényi Dimension:

a) $q = 0$: $\quad d_0 = d_C(\operatorname{supp}\mu)$. (17.49a)

b) $q = 1$: $\quad d_1 := \lim\limits_{q\to 1} d_q = d_I(\mu)$. (17.49b)

c) $q = 2$: $\quad d_2 = d_K$. (17.49c)

5. Lyapunov Dimension

Let $\{\varphi^t\}_{t\in\Gamma}$ be a smooth dynamical system on $M \subset \mathbb{R}^n$ with attractor Λ (or invariant set) and with the invariant ergodic probability measure μ concentrated on Λ. If $\lambda_1 \ge \lambda_2 \ge \dots \ge \lambda_n$ are the Lyapunov exponents with respect to μ and if k is the greatest index for which $\sum\limits_{i=1}^{k} \lambda_i \ge 0$ and $\sum\limits_{i=1}^{k+1} \lambda_i < 0$ hold, then

the value

$$d_L(\mu) = k + \frac{\sum_{i=1}^{k} \lambda_i}{|\lambda_{k+1}|} \tag{17.50}$$

is called the *Lyapunov dimension of the measure* μ.

If $\sum_{i=1}^{n} \lambda_i \geq 0$, then $d_L(\mu) = n$; if $\lambda_1 < 0$, then $d_L(\mu) = 0$.

Ledrappier Theorem: Let $\{\varphi^t\}$ be a discrete system (17.3) on $M \subset \mathbb{R}^n$ with a C^2-function φ and μ, as above, an invariant ergodic probability measure concentrated on the attractor Λ of $\{\varphi^t\}$. Then $d_H(\mu) \leq d_L(\mu)$ holds.

■ **A:** Suppose the attractor $\Lambda \subset \mathbb{R}^2$ of a smooth dynamical system $\{\varphi^t\}$ is covered by N_ε squares with side length ε. Let $\sigma_1 > 1 > \sigma_2$ be the singular values of $D\,\varphi$. Then for the d_B-dimensional volume of the attractor $m_{d_B} \simeq N\varepsilon \cdot \varepsilon^{d_B}$ holds. Every square of side length ε is transformed by φ approximately into a parallelogram with side length $\sigma_2\varepsilon$ and $\sigma_1\varepsilon$. If the covering is made by rhombi with side length $\sigma_2\varepsilon$, then $N_{\sigma_2\varepsilon} \simeq N_\varepsilon \dfrac{\sigma_1}{\sigma_2}$ holds. From the relation $N_\varepsilon \varepsilon^{d_B} \simeq N_{\sigma_2\varepsilon}(\varepsilon\sigma_2)^{d_B}$ one gets directly

$$d_B \simeq 1 - \frac{\ln \sigma_1}{\ln \sigma_2} = 1 + \frac{\lambda_1}{|\lambda_2|}. \tag{17.51}$$

This heuristic calculation gives a hint of the origin of the formula for the Lyapunov dimension.

■ **B:** Suppose the Hénon system (17.6) is given with $a = 1.4$ and $b = 0.3$. The system (17.6) has an attractor Λ (*Hénon attractor*) with a complicated structure for these parameter values. The numerically determined box-counting dimension is $d_B(\Lambda) \simeq 1.26$. It can be shown that there exists an SBR measure for the Hénon attractor Λ. For the Lyapunov exponents λ_1 and λ_2 the formula $\lambda_1 + \lambda_2 = \ln|\det D\,\varphi(x)| = \ln b = \ln 0.3 \simeq -1.204$ holds. With the numerically determined value $\lambda_1 \simeq 0.42$ one gets $\lambda_2 \simeq -1.62$. So,

$$d_L(\mu) \simeq 1 + \frac{0.42}{1.62} \simeq 1.26. \tag{17.52}$$

17.2.4.3 Local Hausdorff Dimension According to Douady and Oesterlé

Let $\{\varphi^t\}_{t\in\Gamma}$ be a smooth dynamical system on $M \subset \mathbb{R}^n$ and Λ a compact invariant set. Suppose that an arbitrary $t_0 \geq 0$ is fixed and let $\Phi := \varphi^{t_0}$.

Theorem of Douady and Oesterlé: Let $\sigma_1(x) \geq \cdots \geq \sigma_n(x)$ be the singular values of $D\Phi(x)$ and let $d \in (0, n]$ be a number written as $d = d_0 + s$ with $d_0 \in \{0, 1, \ldots, n-1\}$ and $s \in [0, 1]$. If $\sup\limits_{x\in\Lambda} [\sigma_1(x)\sigma_2(x)\ldots\sigma_{d_0}(x)\sigma^s_{d_0+1}(x)] < 1$ holds, then $d_H(\Lambda) < d$.

Special Version for Differential Equations: Let $\{\varphi^t\}_{t\in\mathbb{R}}$ be the flow of (17.1), Λ be a compact invariant set and let $\alpha_1(x) \geq \cdots \geq \alpha_n(x)$ be the eigenvalues of the *symmetrized Jacobian matrix* $\dfrac{1}{2}[D\,f(x)^T + D\,f(x)]$ at an arbitrary point $x \in \Lambda$. If $d \in (0, n]$ is a number of the form $d = d_0 + s$ where $d_0 \in \{0, \ldots, n-1\}$ and $s \in [0, 1]$, and $\sup\limits_{x\in\Lambda}[\alpha_1(x) + \cdots + \alpha_{d_0}(x) + s\alpha_{d_0+1}(x)] < 0$ holds, then $d_H(\Lambda) < d$. The quantity

$$d_{D0}(x) = \begin{cases} 0, \text{ if } \alpha_1(x) < 0, \\ \sup\{d\colon 0 \leq d \leq n,\ \alpha_1(x) + \cdots + \alpha_{[d]}(x) + (d - [d])\alpha_{[d]+1}(x) \geq 0\} \text{ otherwise}, \end{cases} \tag{17.53}$$

where $x \in \Lambda$ is arbitrary and $[d]$ is the integer part of d, is called the *Douady-Oesterlé dimension* at the point x. Under the assumptions of the Douady-Oesterlé theorem for differential equations, $d_H(\Lambda) \leq \sup\limits_{x\in\Lambda} d_{DO}(x)$.

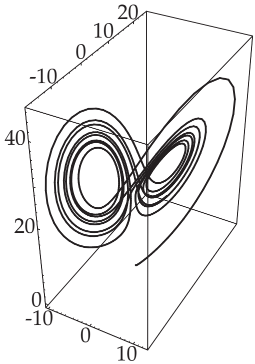


Figure 17.17

■ The Lorenz system (17.2), in the case of $\sigma = 10,\ b = 8/3,\ r = 28$, has an attractor Λ (*Lorenz attractor*) with the numerically determined dimension $d_H(\Lambda) \approx 2.06$ (**Fig. 17.17** is generated by Mathematica). With the Douady-Oesterlé theorem, for arbitrary $b > 1, \sigma > 0$ and $r > 0$ one gets the estimation

$$d_H(\Lambda) \le 3 - \frac{\sigma + b + 1}{\kappa} \quad \text{where} \tag{17.54a}$$

$$\kappa = \frac{1}{2}\left[\sigma + b + \sqrt{(\sigma - b)^2 + \left(\frac{b}{\sqrt{b-1}} + 2\right)\sigma r}\,\right]. \tag{17.54b}$$

17.2.4.4 Examples of Attractors

■ **A:** The *horseshoe mapping* φ occurs in connection with Poincaré mappings containing the transversal intersections of stable and unstable manifolds. The unit square $M = [0,1] \times [0,1]$ is stretched linearly in one coordinate direction and contracted in the other direction. Finally, this rectangle is bent at the middle (**Fig. 17.18**). Repeating this procedure infinitely many times, a sequence of sets $M \supset \varphi(M) \supset \cdots$ is produced, for which

$$\Lambda = \bigcap_{k=0}^{\infty} \varphi^k(M) \tag{17.55}$$

is a compact set and an invariant set with respect to φ. Λ attracts all points of M. Apart from one point, Λ can be described locally as a product "line $\times$ Cantor set".

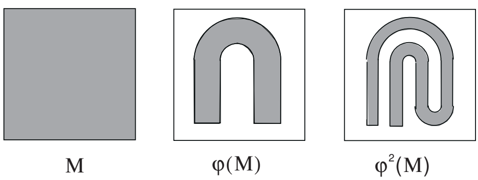


Figure 17.18

■ **B:** Let $\alpha \in (0, 1/2)$ be a parameter and $M = [0,1] \times [0,1]$ be the unit square. The mapping φ: $M \to M$ given by

$$\varphi(x,y) = \begin{cases} (2x, \alpha y), & \text{if } 0 \le x \le \frac{1}{2},\ y \in [0,1], \\ (2x - 1, \alpha y + \frac{1}{2}, & \text{if } \frac{1}{2} < x \le 1,\ y \in [0,1] \end{cases} \tag{17.56a}$$

is called the dissipative *baker's mapping.* Two iterations of the *baker's* mapping are shown in **Fig. 17.19**.

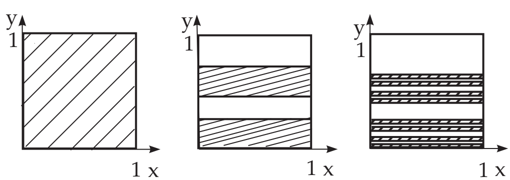


Figure 17.19

The "*flaky pastry structure*" is recognizable. The set

$$\Lambda = \bigcap_{k=0}^{\infty} \varphi^k(M) \tag{17.56b}$$

is invariant under φ and all points from M are attracted by Λ. The value of the Hausdorff dimension is

$$d_H(\Lambda) = 1 + \frac{\ln 2}{-\ln \alpha}. \tag{17.56c}$$

For the dynamical system $\{\varphi^k\}$ there exists an invariant measure μ on M, which is different from the Lebesgue measure. At the points where the derivative exists, the Jacobian matrix is $D\,\varphi^k((x,y)) = \begin{pmatrix} 2^k & 0 \\ 0 & \alpha^k \end{pmatrix}$. Hence, the singular values are $\sigma_1(k,(x,y)) = 2^k$, $\sigma_2(k,(x,y)) = \alpha^k$ and, consequently,

the Lyapunov exponents are $\lambda_1 = \ln 2,\ \lambda_2 = \ln\alpha$ (with respect to the invariant measure μ). For the Lyapunov dimension

$$d_L(\mu) = 1 + \frac{\ln 2}{-\ln\alpha} = d_H(\Lambda) \tag{17.56d}$$

is obtained. Pesin's formula (see 17.2.3,**4.**, (17.40), p. 881) for the metric entropy is valid here, i.e.,

$$h_\mu = \sum_{\lambda_i>0} \lambda_i = \ln 2. \tag{17.56e}$$

■ **C:** Let T be the whole torus with local coordinates (Θ, x, y), as is shown in **Fig. 17.20a**.

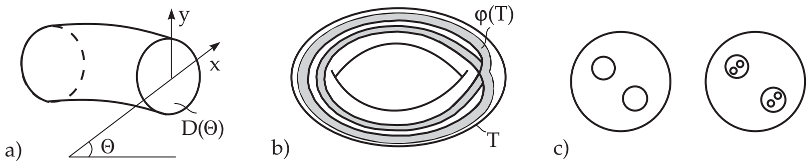


Figure 17.20

Let a mapping $\varphi\colon T \to T$ be defined by

$$\Theta_{k+1} = 2\Theta_k,\ \begin{pmatrix} x_{k+1} \\ y_{k+1} \end{pmatrix} = \frac{1}{2}\begin{pmatrix} \cos\Theta_k \\ \sin\Theta_k \end{pmatrix} + \alpha\begin{pmatrix} x_k \\ y_k \end{pmatrix} \quad (k = 0, 1, \ldots) \tag{17.57}$$

with parameter $\alpha \in (0, 1/2)$. The image $\varphi(T)$, with the intersections $\varphi(T) \cap D(\Theta)$ and $\varphi^2(T) \cap D(\Theta)$, is shown in **Fig. 17.20b** and **Fig. 17.20c**. The result of infinitely many intersections is the set $\Lambda = \bigcap_{k=0}^{\infty} \varphi^k(T)$, which is called a *solenoid*. The attractor Λ consists of a continuum of curves in the length direction, and each of them is dense in Λ, and unstable. The cross-section of the Λ transversal to these curves is a Cantor set.

The Hausdorff dimension is $d_H(\Lambda) = 1 - \frac{\ln 2}{\ln\alpha}$. The set Λ has a neighborhood which is a domain of attraction. Furthermore, the attractor Λ is structurally stable, i.e., the qualitative properties formulated above do not change for C^1-small perturbations of φ.

■ **D:** The solenoid is an example of a *hyperbolic attractor*.

17.2.5 Strange Attractors and Chaos

1. Chaotic Attractor

Let $\{\varphi^t\}_{t\in\Gamma}$ be a dynamical system in the metric space (M, ρ). The attractor Λ of this system is called *chaotic* if there is a *sensitive dependence on the initial condition* in Λ.

The property " sensitive dependence on the initial conditions " will be made more precise in different ways. It is given, e.g., if one of the two following conditions is fulfilled:

a) All motions of $\{\varphi^t\}$ on Λ are unstable in a certain sense.

b) The greatest Lyapunov exponent of $\{\varphi^t\}$ is positive with respect to an invariant ergodic probability measure concentrated on Λ.

■ Sensitive dependence in the sense of **a)** occurs for the solenoid. Property **b)** can be found, e.g., for Hénon attractors.

2. Fractals and Strange Attractors

An attractor Λ of $\{\varphi^t\}_{t\in\Gamma}$ is called *fractal* if it represents neither a finite number of points or a piecewise differentiable curve or surface nor a set which is bounded by some closed piecewise differentiable surface. An attractor is called *strange* if it is chaotic, fractal or both. The notions chaotic, fractal and strange

are used for compact invariant sets analogously even if they are not attractors. A dynamical system is called *chaotic* if it has a compact invariant chaotic set.

■ The mapping

$$x_{n+1} = 2x_n + y_n \pmod 1, \quad y_{n+1} = x_n + y_n \pmod 1 \tag{17.58}$$

(*Anosov diffeomorphism*) is considered on the unit square. The adequate phase space for this system is the torus T^2. It is conservative, has the Lebesgue measure as invariant measure, has a countable number of periodic orbits whose union is dense and is mixing. Otherwise, $\Lambda = T^2$ is an invariant set with integer dimension 2.

3. Systems Chaotic in the Sense of Devaney

Let $\{\varphi^t\}_{t\in\Gamma}$ be a dynamical system in the metric space (M, ρ) with a compact invariant set Λ. The system $\{\varphi^t\}_{t\in\Gamma}$ (or the set Λ) is called *chaotic in the sense* of Devaney, if:

a) $\{\varphi^t\}_{t\in\Gamma}$ is *topologically transitive* on Λ, i.e., there is a positive semiorbit, which is dense in Λ.

b) The periodic orbits of $\{\varphi^t\}_{t\in\Gamma}$ are dense in Λ.

c) $\{\varphi^t\}_{t\in\Gamma}$ is *sensitive with respect to the initial conditions in the sense of* Guckenheimer on Λ, i.e.,

$$\exists\, \varepsilon > 0 \ \forall\, x \in \Lambda \ \forall\, \delta > 0 \quad \exists\, y \in \Lambda \cap U_\delta(x) \quad \exists\, t \geq 0 \ : \rho(\varphi^t(x), \varphi^t(y)) \geq \varepsilon \tag{17.59}$$

where $U_\delta(x) = \{z \in M\colon \rho(x, z) < \delta\}$.

■ Consider the space of the 0–1-sequences

$$\sum = \{s = s_0 s_1 s_2 \ldots, \quad s_i \in \{0, 1\} \quad (i = 0, 1 \ldots)\}.$$

For two sequences $s = s_0 s_1 s_2 \ldots$ and $s' = s'_0 s'_1 s'_2 \ldots$, their distance is defined by

$$\rho(s, s') = \begin{cases} 0, & \text{if } s = s', \\ 2^{-j}, & \text{if } s \neq s', \end{cases}$$

where j is the smallest index for which $s_j \neq s'_j$. So, $(\sum, \rho)$ is a complete metric space which is also compact.

The mapping ρ: $s = s_0 s_1 s_2 \ldots \longmapsto \sigma(s) = s' = s_1 s_2 s_3 \ldots$ is called a *Bernoulli shift*.

The Bernoulli shift is chaotic in the sense of Devaney.

17.2.6 Chaos in One-Dimensional Mappings

For continuous mappings of a compact interval into itself, there exist several sufficient conditions for the existence of chaotic invariant sets. Three examples are to be considered.

Shinai Theorem: Let $\varphi\colon I \to I$ be a continuous mapping of a compact interval I, e.g., $I = [0, 1]$ into itself. Then the system $\{\varphi^k\}$ on I is chaotic in the sense of Devaney if and only if the topological entropy of φ on I, i.e., $h(\varphi)$, is positive.

Sharkovsky Theorem: Consider the following ordering of positive integer numbers:

$$3 \succ 5 \succ 7 \succ \ldots \succ 2\cdot 3 \succ 2\cdot 5 \succ \ldots \succ 2^2\cdot 3 \succ 2^2\cdot 5 \succ \ldots\ldots \succ 2^3 \succ 2^2 \succ 2 \succ 1. \tag{17.60}$$

Let $\varphi\colon I \to I$ be a continuous mapping of a compact interval into itself and suppose $\{\varphi^k\}$ has an n-periodic orbit on I. Then $\{\varphi^k\}$ also has an m-periodic orbit if $n \succ m$.

Block, Guckenheimer and Misiuriewicz Theorem: Let $\varphi : I \to I$ be a continuous mapping of the compact interval I into itself such that $\{\varphi^k\}$ has a $2^n m$-periodic orbit ($m > 1$, odd). Then $h(\varphi) \geq \dfrac{\ln 2}{2^{n+1}}$ holds.

17.2.7 Reconstruction of Dynamics from Time Series

17.2.7.1 Foundations, Reconstruction with Basic Properties

1. Measuring Function, Time Series

Considered is a dynamic system $\{\varphi^t\}_{t\in\Gamma}$ with $\Gamma \in \{\mathbb{Z}_+, \mathbb{R}_+\}$, generated by a mapping $\varphi \in \mathrm{Diff}^1_+(U)$

(see 17.1.4.2, p. 874) or a vector field $f \in \mathrm{X}^1_+(U)$ (see 17.1.4.1. p. 873). A C^1-function $h\colon U \to \mathbb{R}$, the so called *measuring function* is needed to reconstruct the dynamics from measurements. Since in practice only discrete time measurements can be obtained, this will be done in time periods $\{k\tau,\ k = 1, 2, \ldots\} \subset \Gamma$, where $\tau > 0$ is a fixed interval of time. For $m \in \mathbb{N}$ and $\kappa \in \{-1, 1\}$ the time sequence of order m measured with fixed time step $\tau > 0$

$$\left\{\left(h\left(\varphi^{\kappa\tau}(p)\right), h\left(\varphi^{(k+\kappa)\tau}(p)\right), \ldots, h\left(\varphi^{(k+(m-1)\kappa)\tau}(p)\right)\right)\right\}_{k=m-1}^{\infty} \tag{17.61}$$

is called the backward ($\kappa = -1$) or forward ($\kappa = 1$) coordinates of the orbit $\{\varphi^t(p)\}_{t\in\Gamma}$ with $p \in U$ based on the measuring function h.

2. Immersion, Embedding, Theorem of Whitney

Let $U \subset \mathbb{R}^n$ be an open set. The C^1-mapping $\Phi\colon U \to \mathbb{R}^m$ is called an *immersion* if the Jacobian matrix $D\Phi(u)$ has rank n for every $u \in U$. The immersion $\Phi\colon U \to \mathbb{R}^m$ is called *embedding* if Φ is a homeomorphism of U into $\Phi(U)$ ($\Phi(U)$ is considered with the subspace topology of $\mathbb{R}^m$).

Whitney's theorem states that with respect to the open and bounded set $U \subset \mathbb{R}^n$ for every $m \ge 2n+1$ the set of all embeddings $\Phi\colon U \to \mathbb{R}^m$ form an open and dense subset of C^1-mappings $U \to \mathbb{R}^m$. Hence, for $m \ge 2n + 1$ is Φ generically an embedding.

3. Reconstruction Theorem of Takens, Theorem of Kupka and Smale

Considering $\mathrm{Diff}^1_+(U)$ with an arbitrary natural number $m \ge 2n + 1$ the set of all pairs $(\varphi, h) \in \mathrm{Diff}^1_+(U) \times C^1(U, \mathbb{R})$, for which the *reconstruction mapping* (in forward coordinates)

$$p \in U \mapsto \Phi_{\varphi,h}(p) = \left(h(p), h(\varphi^1(p)), \ldots, h(\varphi^{m-1}(p))\right) \tag{17.62}$$

is an embedding, is open and dense in $\mathrm{Diff}^1_+(U) \times C^1(U, \mathbb{R})$. Hence, the property of $\Phi_{\varphi,h}$, to be an embedding is generic for $m \ge 2n + 1$. Such a number m is called *embedding dimension* (see theorem of Takens [17.13]). The theorem of Takens can be applied for differential equations from $\mathrm{X}^1_+(U)$: If $m \ge 2n + 1$ is an arbitrary natural number, then the set of all pairs $(f, h) \in \mathrm{X}^1_+(U) \times C^1(U, \mathbb{R})$ for which the *reconstruction mapping* (in forward coordinates)

$$p \in U \mapsto \Phi_{f,h}(p) = \left(h(p), h(\varphi^1(p)), \ldots, h(\varphi^{m-1}(p))\right) \tag{17.63}$$

(with $\{\varphi^t\}_{t\ge 0}$ as the semi-flow belonging to f) is an embedding, constitutes an open and dense set in $\mathrm{X}^1_+(U) \times C^1(U, \mathbb{R})$. Hence, the property of $\Phi_{\varphi,h}$, to be an embedding is generic.

■ Let the differential equation $\dot{x} = -x \equiv f(x)$ be given in an interval $(-1 - \varepsilon, 1 + \varepsilon)$ ($\varepsilon > 0$, for a sufficiently small ε). Since f is continuously differentiable and $f(-1) = 1 > 0$ and $f(1) = -1$, clearly $f \in \mathrm{X}^1_+(U)$ with $U = (-1, 1)$. This also follows from the explicite solution $\varphi^t(x) = xe^{-t}(t \ge 0, x \in U)$. The theorem of Takens states that for $m \ge 3$ the reconstruction function $\Phi_{f,h}$ is a generic embedding with continuously differentiable measuring function $h : (-1, 1) \to \mathbb{R}$. For example the measuring function $h_1(x) = x$ for the mapping $x \in (-1, 1) \mapsto \Phi_{f,h_1}(x) = (x, xe^{-1}, xe^{-2})$ is obviously an embedding in $\mathbb{R}^3$. However with the measuring function $h\colon (-1, 1) \to \mathbb{R}$ the reconstruction mapping $x \in (-1, 1) \mapsto \Phi_{f,h_2}(x) = (x^2, x^2e^{-2}, x^2e^{-4})$ is not injective and therefore is not an embedding.

The reconstruction theorem of Takens is based on the theorem of Kupka and Smale: The set of all diffeomorphisms $\varphi \in \mathrm{Diff}^1_+(U)$, such that their period points are hyperbolic and $W^s(p)$ is transversal to $W^u(q)$ for arbitrary period point, forms a set of second Baire category, that is, such diffeomorphisms are typical in $\mathrm{Diff}^1_+(U)$. The condition $m \ge 2n + 1$ follows from the fact that for typical $(\varphi, h) \in \mathrm{Diff}^1_+(U) \times C^1(U, \mathbb{R})$ mapping $\Phi_{\varphi,h}$ is an immersion in the neighborhood of the period points, therefore it can be extended into an embedding on the entire U.

4. Dynamic in the Reconstruction Space

The theorem of Takens implies that for a generic $(\varphi, h) \in \mathrm{Diff}^1_+(U) \times C^1(U, \mathbb{R})$ the set $\Phi(U)$ (reconstruction space) with $\Phi = \Phi_{\varphi,h}$ is an immerse and homeomorphic image of U and the mapping

$\psi = \Phi \circ \varphi \circ \Phi^{-1}$ is defined on $\Phi(U)$. The topologic properties of the stationary points and periodic orbits of the (unknown) system $\{\varphi^k\}_{k\in\mathbb{Z}_+}$ defined on U and the system $\{\psi^k\}_{k\in\mathbb{Z}_+}$ defined on $\Phi(U)$ are identical to those of the eigenvalues of the Jacobi matrices. Similarly the entropies and dimensions, e.g. correlation dimension (see 17.2.7.2, p. 891), correspond to the Lyapunov exponents associated with invariant measure. The mapping ψ on $\Phi(u)$ is completely described in all points being given in the time series. For example select $\tau = 1$. Let point $x_k = \left(h(\varphi^k(p)), \ldots, h(\varphi^{k+m-1}(p))\right) \in \mathbb{R}^m$, $k \in \mathbb{Z}_+$ be given. Clearly $x_k = \Phi(q_k)$, with $q_k = \varphi^k(p)$. Then $\psi(x_k) = (\Phi \circ \varphi \circ \Phi^{-1})(\Phi(q_k)) = x_{k+1}$, i.e., $\psi\,(h(q_k), \ldots, h(q_{k+m-1})) = (h(q_{k+1}), h(q_{k+2}), \ldots, h(q_{k+m}))$. From the measurements of the orbits (with $\Gamma = \mathbb{Z}_+$) of ψ defined on $\Phi(U)$ the entire dynamics of φ defined on U can be obtained.

17.2.7.2 Reconstructions with Prevalent Properties

1. Prevalence as a Generic Metric

Prevalence or generic metric is an extension to infinite dimension of the wellknown concept "almost everywhere in Lebesgue-measure" (see 12.9.1,**2.**, p. 694) known in finite dimensions, so it differs from the corresponding notions for the sets of the second Baire category. A Borel set S in a Banach space B is prevalent (see [17.13]) if there is a finite Borel measure μ with support K (see 17.2.1.1,**1.**, p. 876), such that $\mu(S + x) = \mu(S) = \mu(K)$ for all $x \in B$.

■**A:** Every Borel set of a finite dimensional vector space with measure zero complement is prevalent.

■**B:** The union and intersection of finitely many prevalent sets are also prevalent.

■**C:** Let $C^k(\overline{U}, \mathbb{R})$, $U \subset \mathbb{R}^n$ be the Banach space of all scalar valued functions from $C^k(\overline{U}, \mathbb{R})$ whose partial derivatives up to order k are continuous on $\overline{U}$. If $U \in \mathbb{R}^n$ is open and connected, then the prevalent subsets of $C^k(U, \mathbb{R})$, are dense in this space.

2. Reconstruction Theorems of Sauer, Yorke and Casdagli

Let $\{\varphi^t\}_{t\geq 0}$ be a continuous dynamics generated by the vector field $f \in X^1_+$ (see 17.1.4.1, p. 873) and let A be a compact subset in U with fractal dimension $\overline{d}_C(A) = d$. Furthermore let $m > 2d$ be an integer and $\tau > 0$ arbitrary. The process $\{\varphi^t\}_{t\geq 0}$ restricted to A might have only finitely many stationary points, there is no orbit with period τ or 2τ, and only finitely many orbits with periods $3\tau, 4\tau, \ldots, m\tau$ where the multiplyers of these periodic orbits (except by 1) are all different. The set of all measuring functions $h : U \to \mathbb{R}$ is a prevalent set $C^1(\overline{U}, \mathbb{R})$ if the reconstruction function $\Phi_{f,h,\tau}$

$$p \in U \mapsto \Phi_{f,h,\tau}(p) = \left(h\left(\varphi^{(m-1)\tau}(p)\right), h\left(\varphi^{(m-2)\tau}(p)\right), \ldots, h(p)\right) \qquad (17.64)$$

satisfies the following conditions:

a) $\Phi_{f,h,\tau}$ is injective on A;

b) $\Phi_{f,h,\tau}$ is an immersion on every subset $\tilde{U} \subset A$, such that $\tilde{U} = \Psi(W)$, whith $W = G \cap \mathbb{R}^k \times \underbrace{\{0, \ldots, 0\}}_{n-k}$, $G \subset \mathbb{R}^n$ open, $\Psi\colon G \to \mathbb{R}^n$ is a C^1-mapping and $k \leq d$. (Theorem of Sauer, Yorke, Casdagli see [17.13].)

3. Estimation of Correlation Dimension

Given the dynamic system $\{\varphi^t\}_{t\in\Gamma}$ mit $\Gamma \in \{\mathbb{Z}_+, \mathbb{R}_+\}$, generated either by $\varphi \in \mathrm{Diff}^1_+(U)$ or $f \in X^1_+(U)$. Let $\{\varphi^t\}_{t\in\Gamma}$ have an attractor Λ in U with invariant probability measure μ. Let $h : U \to \mathbb{R}$ be a measuring function, $m \in \mathbb{N}$ the order parameter, $\tau = 1$ the time stepsize and for $i = 1, 2, \ldots$

$$x_i = (y_i, y_{i+1}, \ldots, y_{i+m}) \in \mathbb{R}^{m+1} \qquad (17.65)$$

with $y_i = h(\varphi^i(p))$ being a forward coordinate of the $m+1$ order time series of the orbit $\{\varphi^t\}_{t\in\Gamma}$ with $p \in U$. The distance of vectors x_i and x_j is defined as $\mathrm{dist}(x_i, x_j) = \max\limits_{0\leq s\leq m} |y_{i+s} - y_{j+s}|$. Let $N > m$

denote a natural number and $\varepsilon > 0$ a real number, then the expression

$$C^m(\varphi) = \limsup_{N\to\infty} \frac{1}{N^2}\{\text{number of ordered pairs } (x_i, x_j)\colon i, j \in \{1, \ldots, N\}, \operatorname{dist}(x_i, x_j) < \varepsilon\} \quad (17.66)$$

is called (discrete) *correlation integral* (with respect to m and ε). If the $d_K(m) = \lim_{\varepsilon\to 0} \frac{\ln C^m(\varepsilon)}{\ln \varepsilon}$ exists, then it is an estimation of the correlation dimension d_K. The Takens theorem implies that $h \in C^1(\overline{U}, \mathbb{R})$ is generic for $m \geq 2n$, and the theorem of Sauer, Yorke, Casdagli implies that it is prevalent for $m+1 \geq \overline{d}_C(\Lambda)$ with backward coordinates.

■ The Lorenz system (17.2) (see 17.1.1.1,**2.**, p.858) belongs to $X^1_+(U)$, where U can be selected as $U := \{(x, y, z) \in \mathbb{R}^3 : \frac{1}{2}[x^2 + y^2 + (z - \sigma - r)^2] < c\}$ ($c > 0$, sufficiently large). Clearly the Lorenz attractor Λ (with $\sigma = 10$, $b = 8/3$, $r = 28$) is in U. The theorem of Douady-Oesterlé (see 17.2.4.3, p. 886) gives the upperbound $d_H(\Lambda) \leq 2.421$. Numerical integration with the Box-Counting-Method gives $d_H(\Lambda) \approx 2.06$. Estimation of the correlation dimension in natural measure using the embedding method with time series in backward coordinates ($\tau \approx 0.12$) gives the value $d_K \approx 2.03$ (Grassberger, [17.12]) for the Lorenz attractor .

17.3 Bifurcation Theory and Routes to Chaos

17.3.1 Bifurcations in Morse-Smale Systems

Let $\{\varphi^t_\varepsilon\}_{t\in\Gamma}$ be a dynamical system generated by a differential equation or by a mapping on $M \subset \mathbb{R}^n$, which additionally depends on a parameter $\varepsilon \in V \subset \mathbb{R}^l$. Every change of the topological structure of the phase portrait of the dynamical system for small changes of the parameter is called *bifurcation*. The parameter $\varepsilon = 0 \in V$ is called the *bifurcation value* if there exist parameter values $\varepsilon \in V$ in every neighborhood of 0 such that the dynamical systems $\{\varphi^t_\varepsilon\}$ and $\{\varphi^t_0\}$ are not topologically equivalent or conjugated on M. The smallest dimension of the parameter space for which a bifurcation can be observed is called the *codimension* of the bifurcation.

One distinguishes between *local* bifurcations, which occur in the neighborhood of a single orbit of the dynamical system, and *global* bifurcations, which affect a large part of the phase space.

17.3.1.1 Local Bifurcations in Neighborhoods of Steady States

1. Center Manifold Theorem

Consider a parameter-dependent differential equation

$$\dot{x} = f(x, \varepsilon) \quad \text{or} \quad \dot{x}_i = f_i(x_1, \ldots, x_n, \varepsilon_1, \ldots, \varepsilon_l) \quad (i = 1, 2, \ldots, n) \qquad (17.67)$$

with $f\colon M \times V \to \mathbb{R}^n$, where $M \subset \mathbb{R}^n$ and $V \subset \mathbb{R}^l$ are open sets and f is supposed to be r times continuously differentiable. Equation (17.67) can be interpreted as a parameter-free differential equation $\dot{x} = f(x, \varepsilon)$, $\dot{\varepsilon} = 0$ in the phase space $M \times V$. From the Picard-Lindelöf theorem and the theorem on differentiability with respect to the initial values (see 17.1.1.1, **2.**, p. 857) it follows that (17.67) has a locally unique solution $\varphi(\cdot, p, \varepsilon)$ with initial point p at time $t = 0$ for arbitrary $p \in M$ and $\varepsilon \in V$, which is r times continuously differentiable with respect to p and ε. Suppose all solutions exist on the whole of $\mathbb{R}$.

Furthermore, it is supposed that the system (17.67) has the equilibrium point $x = 0$ at $\varepsilon = 0$, i.e., $f(0, 0) = 0$ holds. Let $\lambda_1, \ldots, \lambda_s$ be the eigenvalues of $D_x f(0, 0) = \left[\frac{\partial f_i}{\partial x_j}(0, 0)\right]^n_{i,j=1}$ with $\operatorname{Re}\lambda_j = 0$.

Furthermore, suppose, $D_x f(0, 0)$ has exactly m eigenvalues with negative real part and $k = n - s - m$ eigenvalues with positive real part.

According to the *center manifold theorem* for differential equations (*theorem of Shoshitaishvili*, see [17.12]), the differential equation (17.67), for ε with a sufficiently small norm $\|\varepsilon\|$ in the neighborhood

of 0, is topologically equivalent to the system

$$\dot{x} = F(x, \varepsilon) \equiv Ax + g(x, \varepsilon), \quad \dot{y} = -y, \quad \dot{z} = z \tag{17.68}$$

with $x \in \mathbb{R}^s$, $y \in \mathbb{R}^m$ and $z \in \mathbb{R}^k$, where A is a matrix of type (s, s) with eigenvalues $\lambda_1, \dots, \lambda_s$, and g represents a C^r-function with $g(0, 0) = 0$ and $D_x g(0, 0) = 0$.
It follows from representation (17.68) that the bifurcations of (17.67) in a neighborhood of 0 are uniquely described by the differential equation

$$\dot{x} = F(x, \varepsilon). \tag{17.69}$$

Equation (17.69) represents the *reduced differential equation* to the *local center manifold* $W^c_{loc} = \{x, y, z\colon y = 0, z = 0\}$ of (17.68). The reduced differential equation (17.69) can often be transformed into a relatively simple form, e.g., with polynomials on the right-hand side, by a non-linear parameter-dependent coordinate transformation so that the topological structure of its phase portrait does not change close to the considered equilibrium point. This form is a so-called *normal form.* A normal form cannot be determined uniquely; in general, a bifurcation can be described equivalently by different normal forms.

2. Saddle-Node Bifurcation and Transcritical Bifurcation

Suppose (17.67) is given with $l = 1$, where f is continuously differentiable at least twice and $D_x f(0, 0)$ has the eigenvalue $\lambda_1 = 0$ and $n - 1$ eigenvalues λ_j with $\mathrm{Re}\lambda_j \neq 0$. According to the center manifold theorem, in this case, all bifurcations (17.67) near 0 are described by a one-dimensional reduced differential equation (17.69). Obviously, here $F(0, 0) = \frac{\partial F}{\partial x}(0, 0) = 0$. If, additionally, it is supposed that $\frac{\partial^2}{\partial x^2} F(0, 0) \neq 0$, $\frac{\partial F}{\partial \varepsilon}(0, 0) \neq 0$ and the right-hand side of (17.69) is expanded according to the Taylor formula, then this representation can be transformed by coordinate transformation (see [17.6]) into the normal form

$$\dot{x} = \alpha + x^2 + \cdots \tag{17.70}$$

$\left(\text{for } \frac{\partial^2 F}{\partial x^2}(0, 0) > 0\right)$ or $\dot{x} = \alpha - x^2 + \cdots \quad \left(\text{for } \frac{\partial^2 F}{\partial x^2}(0, 0) < 0\right)$, where $\alpha = \alpha(\varepsilon)$ is a differentiable function with $\alpha(0) = 0$ and the points indicate higher-order terms. For $\alpha < 0$, (17.70) has two equilibrium points close to $x = 0$, among which one is stable, the other is unstable. For $\alpha = 0$, these equilibrium points fuse into $x = 0$, which is unstable. For $\alpha > 0$, (17.70) has no equilibrium point near to 0 **(Fig. 17.21b)**.
The multidimensional case results in a *saddle-node bifurcation* in a neighborhood of 0 in (17.67). This bifurcation is represented in **Fig. 17.22** for $n = 2$ and $\lambda_1 = 0, \lambda_2 < 0$. The representation of the saddle-node bifurcation in the extended phase space is shown in **Fig. 17.21a**. For sufficiently smooth vector fields (17.67), the saddle-node bifurcations are generic.

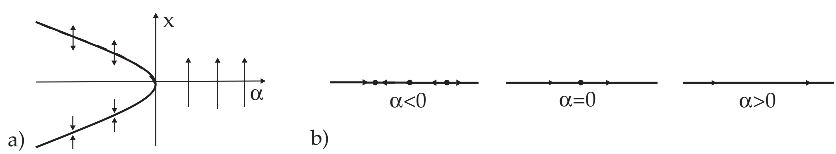


Figure 17.21

If among the conditions which yield a saddle-node bifurcation for F, the condition $\frac{\partial F}{\partial \varepsilon}(0, 0) \neq 0$ is replaced by the conditions $\frac{\partial F}{\partial \varepsilon}(0, 0) = 0$ and $\frac{\partial^2 F}{\partial x \partial \varepsilon}(0, 0) \neq 0$, then one gets from (17.69) the truncated

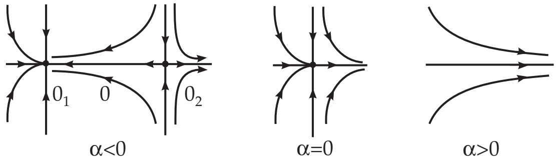


Figure 17.22

normal form (without higher-order terms) $\dot{x} = \alpha x - x^2$ of a *transcritical bifurcation*. For $n = 2$ and $\lambda_2 < 0$, the transcritical bifurcation together with the bifurcation diagram is shown in **Fig. 17.23**. Saddle-node and transcritical bifurcations have codimension 1-bifurcations.

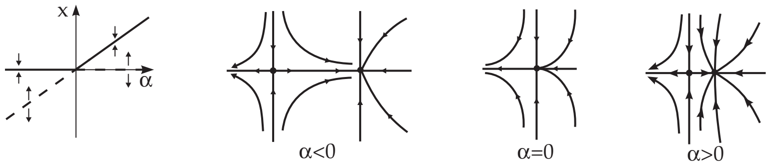


Figure 17.23

3. Hopf Bifurcation

Consider (17.67) with $n \geq 2$, $l = 1$ and $r \geq 4$. Suppose that $f(0, \varepsilon) = 0$ is valid for all ε with $|\varepsilon| \leq \varepsilon_0$ ($\varepsilon_0 > 0$ sufficiently small). Suppose the Jacobian matrix $D_x f(0, 0)$ has the eigenvalues $\lambda_1 = \overline{\lambda_2} = \mathrm{i}\omega$ with $\omega \neq 0$ and $n - 2$ eigenvalues λ_j with $\mathrm{Re}\lambda_j \neq 0$. According to the center manifold theorem, the bifurcation is described by a two-dimensional reduced differential equation (17.69) of the form

$$\dot{x} = \alpha(\varepsilon)x - \omega(\varepsilon)y + g_1(x, y, \varepsilon), \quad \dot{y} = \omega(\varepsilon)x + \alpha(\varepsilon)y + g_2(x, y, \varepsilon) \tag{17.71}$$

where α, ω, g_1 and g_2 are differentiable functions and $\omega(0) = \omega$ and also $\alpha(0) = 0$ hold. By a non-linear complex coordinate transformation and by the introduction of polar coordinates (r, ϑ), (17.71) can be written in the normal form

$$\dot{r} = \alpha(\varepsilon)r + a(\varepsilon)r^3 + \cdots, \quad \dot{\vartheta} = \omega(\varepsilon) + b(\varepsilon)r^2 + \cdots \tag{17.72}$$

where dots denote the terms of higher order. The Taylor expansion of the coefficient functions of (17.72) yields the truncated normal form

$$\dot{r} = \alpha'(0)\varepsilon r + a(0)r^3, \quad \dot{\vartheta} = \omega(0) + \omega'(0)\varepsilon + b(0)r^2. \tag{17.73}$$

The theorem of Andronov and Hopf guarantees that (17.73) describes the bifurcations of (17.72) in a neighborhood of the equilibrium point for $\varepsilon = 0$.

The following cases occur for (17.73) under the assumption $\alpha'(0) > 0$:

1. $a(0) < 0$ **(Fig. 17.24a)**. :

a) $\varepsilon > 0$: Stable limit cycle and unstable equilibrium point.

b) $\varepsilon = 0$: Cycle and equilibrium point fuse into a stable equilibrium point.

c) $\varepsilon < 0$: All orbits close to $(0, 0)$ tend as in b) for $t \to +\infty$ spirally to the equilibrium point $(0, 0)$.

2. $a(0) > 0$ **(Fig. 17.24b)** :

a) $\varepsilon < 0$: Unstable limit cycle.

b) $\varepsilon = 0$: Cycle and equilibrium point fuse into an unstable equilibrium point.

c) $\varepsilon > 0$: Spiral type unstable equilibrium point as in b).

The interpretation of the above cases for the initial system (17.67) shows the bifurcation of a limit cycle of a compound equilibrium point (*compound focus point of multiplicity* 1), which is called a *Hopf bifur-*



Figure 17.24

cation (or *Andronov-Hopf bifurcation*). The case $a(0) < 0$ is also called *supercritical*, the case $a(0) > 0$ *subcritical* (supposing that $\alpha'(0) > 0$). The case $n = 3, \lambda_1 = \overline{\lambda_2} = \mathrm{i},\ \lambda_3 < 0, \alpha'(0) > 0$ and $a(0) < 0$ is shown in **Fig. 17.25**.

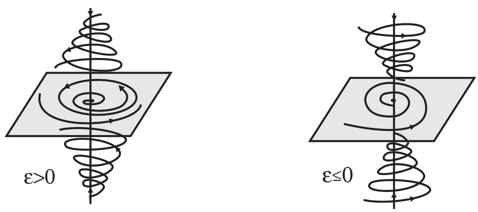


Figure 17.25

Hopf bifurcations are generic and have codimension 1. The cases above illustrate the fact that a supercritical Hopf bifurcation under the above assumptions can be recognized by the stability of a focus: Suppose the eigenvalues $\lambda_1(\varepsilon)$ and $\lambda_2(\varepsilon)$ of the Jacobian matrix on the right-hand side of (17.67) at 0 are pure imaginary for $\varepsilon = 0$, and for the other eigenvalues λ_j $\mathrm{Re}\lambda_j \neq 0$ holds. Suppose furthermore that $\dfrac{d}{d\varepsilon}\,\mathrm{Re}\lambda_1(\varepsilon)_{|\varepsilon=0} > 0$ and let 0 be an asymptotically stable focus for (17.67) at $\varepsilon = 0$. Then there is a supercritical Hopf bifurcation in (17.67) at $\varepsilon = 0$.

■ The van der Pol differential equation $\ddot{x} + \varepsilon(x^2 - 1)\dot{x} + x = 0$ with parameter ε can be written as a planar differential equation

$$\dot{x} = y, \quad \dot{y} = -\varepsilon(x^2 - 1)y - x. \tag{17.74}$$

For $\varepsilon = 0$, (17.74) becomes the harmonic oscillator equation and it has only periodic solutions and an equilibrium point, which is stable but not asymptotically stable. With the transformation $u = \sqrt{\varepsilon}\,x,\ v = \sqrt{\varepsilon}\,y$ for $\varepsilon > 0$ (17.74) is transformed into the planar differential equation

$$\dot{u} = v, \quad \dot{v} = -u - (u^2 - \varepsilon)\,v. \tag{17.75}$$

For the eigenvalues of the Jacobian matrix at the equilibrium point $(0,0)$ of (17.75):

$\lambda_{1,2}(\varepsilon) = \dfrac{\varepsilon}{2} \pm \sqrt{\dfrac{\varepsilon^2}{4} - 1}$ and so $\lambda_{1,2}(0) = \pm\mathrm{i}$ and $\dfrac{d}{d\varepsilon}\,\mathrm{Re}\lambda_1(\varepsilon)_{|\varepsilon=0} = \dfrac{1}{2} > 0$.

As shown in the example of 17.1.2.3, **1.**, p. 863, $(0,0)$ is an asymptotically stable equilibrium point of (17.75) for $\varepsilon = 0$. There is a supercritical Hopf bifurcation for $\varepsilon = 0$, and for small $\varepsilon > 0$, $(0,0)$ is an unstable focus surrounded by a limit cycle whose amplitude is increasing as ε increases.

4. Bifurcations in Two-Parameter Differential Equations

1. Cusp Bifurcation Suppose the differential equation (17.67) is given with $r \geq 4$ and $l = 2$. Let the Jacobian matrix $D_x f(0,0)$ have the eigenvalue $\lambda_1 = 0$ and $n-1$ eigenvalues λ_j with $\mathrm{Re}\lambda_j \neq 0$ and suppose that for the reduced differential equation (17.69) $F(0,0) = \dfrac{\partial F}{\partial x}(0,0) = \dfrac{\partial^2 F}{\partial x^2}(0,0) = 0$ und

$l_3 := \dfrac{\partial^3 F}{\partial x^3}(0,0) \neq 0$. Then the Taylor expansion of F close to $(0,0)$ leads to the truncated normal form (without higher-order terms see, [17.1])

$$\dot{x} = \alpha_1 + \alpha_2 x + \operatorname{sign} l_3 x^3 \tag{17.76}$$

with the parameters α_1 and α_2. The set $\{(\alpha_1, \alpha_2, x)\colon \alpha_1 + \alpha_2 x + \operatorname{sign} l_3 x^3 = 0\}$ represents a surface in extended phase space and this surface is called a *cusp* **(Fig. 17.26a)**.

In the following, it is supposed $l_3 < 0$. The non-hyperbolic equilibrium points of (17.76) are defined by the system $\alpha_1 + \alpha_2 x - x^3 = 0$, $\alpha_2 - 3x^2 = 0$ and thus they lie on the curves S_1 and S_2, which are determined by the set $\{(\alpha_1, \alpha_2)\colon 27\alpha_1^2 - 4\alpha_2^3 = 0\}$ and form a *cusp* **(Fig. 17.26b)**. If $(\alpha_1, \alpha_2) = (0,0)$ then the equilibrium point 0 of (17.76) is stable. The phase portrait of (17.67) in a neighborhood of 0, e.g., for $n = 2$, $l_3 < 0$ and $\lambda_1 = 0$ is shown in **Fig. 17.26c** for $\lambda_2 < 0$ (*triple node*) and in **Fig. 17.26d** for $\lambda_2 > 0$ (*triple saddle*) (see [17.6]).

At transition from $(\alpha_1, \alpha_2) = (0,0)$ into the interior of domain 1 **(Fig. 17.26b)** the compound node-type non-hyperbolic equilibrium point 0 of (17.67) splits into three hyperbolic equilibrium points (two stable nodes and a saddle) (*supercritical pitchfork bifurcation*).

In the case of the two-dimensional phase space of (17.67) the phase portraits are shown in **Fig. 17.26c,e**. When the parameter pair of $S_i \setminus \{(0,0)\}$ $(i = 1, 2)$ traverse from 1 into 2 then a double saddle node-type equilibrium point is formed which finally vanishes. A stable hyperbolic equilibrium point remains.

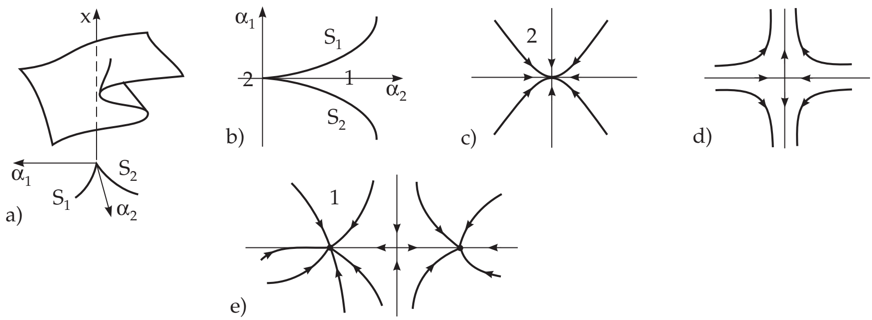


Figure 17.26

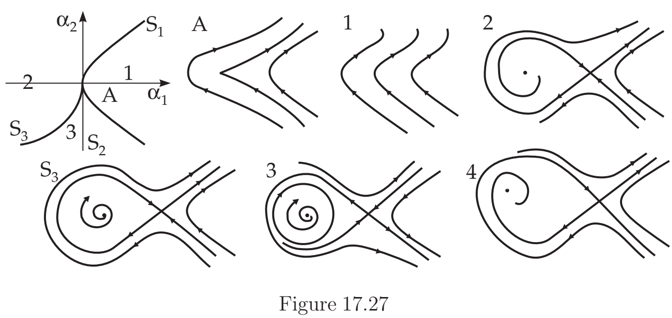


Figure 17.27

2. Bogdanov-Takens Bifurcation Suppose, for (17.67), $n \geq 2$, $l = 2$, $r \geq 2$ hold and the matrix $D_x f(0,0)$ has two eigenvalues $\lambda_1 = \lambda_2 = 0$ and $n - 2$ eigenvalues λ_j with $\mathrm{Re}\lambda_j \neq 0$. Let the reduced

two-dimensional differential equation (17.69) be topologically equivalent to the planar system

$$\dot{x} = y, \quad \dot{y} = \alpha_1 + \alpha_2 x + x^2 - xy. \tag{17.77}$$

Then there is a saddle-node bifurcation on the curve $S_1 = \{(\alpha_1, \alpha_2)\colon \alpha_2^2 - 4\alpha_1 = 0\}$. At the transition on the curve $S_2 = \{(\alpha_1, \alpha_2)\colon \alpha_1 = 0, \alpha_2 < 0\}$ from the domain $\alpha_1 < 0$ into the domain $\alpha_1 > 0$ a stable limit cycle arises by a Hopf bifurcation and on the curve $S_3 = \{(\alpha_1, \alpha_2)\colon \alpha_1 = -k\alpha_2^2 + \cdots\}$ $(k > 0$, constant) there exists a separatrix loop for the original system **(Fig. 17.27)**, which bifurcates into a stable limit cycle in domain 3 (see [17.6], [17.9]).

This bifurcation is of a global nature and one says that *a single periodic orbit arises from the homoclinic orbit of a saddle* or *a separatrix loop* vanishes.

3. Generalized Hopf Bifurcation Suppose that the assumptions of the Hopf bifurcation with $r \geq 6$ are fulfilled for (17.67), and the two-dimensional reduced differential equation after a coordinate transformation into polar coordinates has the normal form $\dot{r} = \varepsilon_1 r + \varepsilon_2 r^3 - r^5 + \cdots$, $\dot{\vartheta} = 1 + \cdots$. The bifurcation diagram **(Fig. 17.28)** of this system contains the line $S_1 = \{(\varepsilon_1, \varepsilon_2)\colon \varepsilon_1 = 0,\ \varepsilon_2 \neq 0\}$, whose points represent a Hopf bifurcation (see [17.4], [17.6]). There exist two periodic orbits in domain 3, among which one is stable, the other one is unstable. On the curve $S_2 = \{(\varepsilon_1, \varepsilon_2)\colon \varepsilon_2^2 + 4\varepsilon_2 > 0,\ \varepsilon_1 < 0\}$, these non-hyperbolic cycles fuse into a compound cycle which disappears in domain 2.

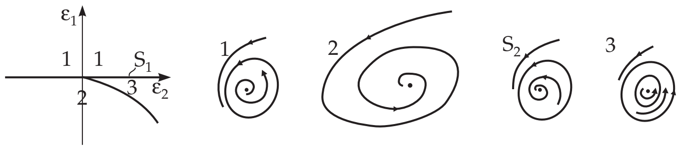


Figure 17.28

5. Breaking Symmetry

Some differential equations (17.67) have *symmetries* in the following sense: There exists a linear transformation T (or a group of transformations) such that $f(Tx, \varepsilon) = T\,f(x, \varepsilon)$ holds for all $x \in M$ and $\varepsilon \in V$. An orbit γ of (17.67) is called *symmetric with respect to* T if $T\gamma = \gamma$.

One talks about a *symmetry breaking* bifurcation at $\varepsilon = 0$, e.g., in (17.67) (for $l = 1$), if there is a stable equilibrium point or a stable limit cycle for $\varepsilon < 0$, which is always symmetric with respect to T, and for $\varepsilon = 0$ two further stable steady states or limit cycles arise, which are no-longer symmetric with respect to T.

■ For system (17.67) with $f(x, \varepsilon) = \varepsilon x - x^3$ the transformation T defined as $T\colon x \mapsto -x$ is a symmetry, since $f(-x, \varepsilon) = -f(x, \varepsilon)$ $(x \in \mathbb{R}, \varepsilon \in \mathbb{R})$. For $\varepsilon < 0$ the point $x_1 = 0$ is a stable equilibrium point. For $\varepsilon > 0$, besides $x_1 = 0$, there exist the two other equilibrium points $x_{2,3} = \pm\sqrt{\varepsilon}$; both are non-symmetric.

17.3.1.2 Local Bifurcations in a Neighborhood of a Periodic Orbit

1. Center Manifold Theorem for Mappings

Let γ be periodic orbit of (17.67) for $\varepsilon = 0$ with multipliers $\rho_1, \ldots, \rho_{n-1}, \rho_n = 1$. A bifurcation close to γ is possible, if when changing ε, at least one of the multipliers lies on the complex unit circle. The use of a surface transversal to γ leads to the parameter-dependent Poincaré mapping

$$x \longmapsto P(x, \varepsilon). \tag{17.78}$$

Then, with open sets $E \subset \mathbb{R}^{n-1}$ and $V \subset \mathbb{R}^l$ let $P\colon E \times V \to \mathbb{R}^{n-1}$ be a C^r-mapping where the mapping $\tilde{P}\colon E \times V \to \mathbb{R}^{n-1} \times \mathbb{R}^l$ with $\tilde{P}(x, \varepsilon) = (P(x, \varepsilon), \varepsilon)$ should be a C^r-diffeomorphism. Furthermore, let $P(0, 0) = 0$ and suppose the Jacobian matrix $D_x P(0, 0)$ has s eigenvalues $\rho_1, \ldots, \rho_s$ with $|\rho_i| = 1$, m eigenvalues $\rho_{s+1}, \ldots, \rho_{s+m}$ with $|\rho_i| < 1$ and $k = n - s - m - 1$ eigenvalues $\rho_{s+m+1}, \ldots, \rho_{n-1}$

with $|\rho_i| > 1$. Then, according to the *the center manifold theorem for mappings* (see [17.4]), close to $(0,0) \in E \times V$, the mapping $\tilde{P}$ is topologically conjugate to the mapping

$$(x, y, z, \varepsilon) \longmapsto (F(x, \varepsilon), A^s y, A^u z, \varepsilon) \tag{17.79}$$

near $(0,0) \in \mathbb{R}^{n-1} \times \mathbb{R}^l$ with $F(x, \varepsilon) = A^c x + g(x, \varepsilon)$. Here g is a C^r-differentiable mapping satisfying the relations $g(0,0) = 0$ and $D_x g(0,0) = 0$. The matrices A^c, A^s and A^u are of type (s,s), (m,m) and (k,k), respectively.

It follows from (17.79) that bifurcations of (17.78) close to $(0,0)$ are described only by the *reduced mapping*

$$x \longmapsto F(x, \varepsilon) \tag{17.80}$$

on the *local center manifold* $W^c_{loc} = \{(x, y, z)\colon\ y = 0,\ z = 0\}$.

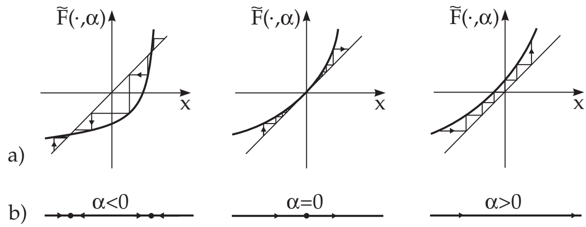


Figure 17.29

2. Bifurcation of Double Semistable Periodic Orbits

Let the system (17.67) be given with $n \geq 2$, $r \geq 3$ and $l = 1$. Suppose, at $\varepsilon = 0$, the system (17.67) has periodic orbit γ with multipliers $\rho_1 = +1$, $|\rho_i| \neq 1$ $(i = 2, 3, \ldots, n-1)$ and $\rho_n = 1$. According to the center manifold theorem for mappings, the bifurcations of the Poincaré mapping (17.78) are described by the one-dimensional reduced mapping (17.80) with $A^c = 1$. If $\dfrac{\partial^2 F}{\partial x^2}(0,0) \neq 0$ and $\dfrac{\partial F}{\partial \varepsilon}(0,0) \neq 0$ is supposed, then it leads to the normal forms

$$x \longmapsto \tilde{F}(x, \alpha) = \alpha + x + x^2 \quad \left(\text{for } \frac{\partial^2 F}{\partial x^2}(0,0) > 0\right) \quad \text{or} \tag{17.81a}$$

$$x \longmapsto \alpha + x - x^2 \quad \left(\text{for } \frac{\partial^2 F}{\partial x^2}(0,0) < 0\right). \tag{17.81b}$$

The iterations from (17.81a) close to 0 and the corresponding phase portraits are represented in **Fig. 17.29a** and in **Fig. 17.29b** for different α (see [17.6]). Close to $x = 0$ there are for $\alpha < 0$ a stable and an unstable equilibrium point, which fuse for $\alpha = 0$ into the unstable steady state $x = 0$. For $\alpha > 0$ there exists no equilibrium point close to $x = 0$. The bifurcation described by (17.81a) in (17.80) is called a *subcritical saddle node bifurcation for mappings*.

In the case of the differential equation (17.67), the properties of the mapping (17.81a) describe the *bifurcation of a double semistable periodic orbit*: For $\alpha < 0$ there exists a stable periodic orbit γ_1 and an unstable periodic orbit γ_2, which fuse for $\alpha = 0$ into a semistable orbit γ, which disappears for $\alpha > 0$ **(Fig. 17.30a,b)**.

3. Period Doubling or Flip Bifurcation

Let system (17.67) be given with $n \geq 2$, $r \geq 4$ and $l = 1$. Considered is a periodic orbit γ of (17.67) at $\varepsilon = 0$ with the multipliers $\rho_1 = -1$, $|\rho_i| \neq 1$ $(i = 2, \ldots, n-1)$, and $\rho_n = 1$. The bifurcation behavior

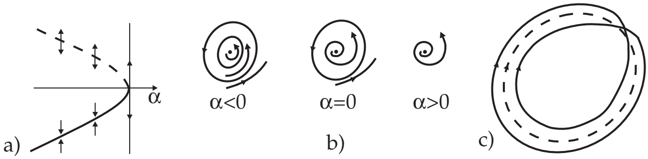


Figure 17.30

of the Poincaré mapping in the neighborhood of 0 is described by the one-dimensional mapping (17.80) with $A^c = -1$, if supposing the normal form

$$x \longmapsto \tilde{F}(x,\alpha) = (-1+\alpha)x + x^3. \tag{17.82}$$

The steady state $x = 0$ of (17.82) is stable for small $\alpha \geq 0$ and unstable for $\alpha < 0$. The second iterated mapping $\tilde{F}^2$ has for $\alpha < 0$ two further stable fixed points besides $x = 0$ for $x_{1,2} = \pm\sqrt{-\alpha} + o(|\alpha|)$, which are not fixed points of $\tilde{F}$. Consequently, they must be points of period 2 of (17.82).

In general, for a C^4-mapping (17.80) there is a two-periodic orbit at $\varepsilon = 0$, if the following conditions are fulfilled (see [17.2]):

$$\begin{aligned} &F(0,0) = 0, && \frac{\partial F}{\partial x}(0,0) = -1, && \frac{\partial F^2}{\partial \varepsilon}(0,0) = 0, \\ &\frac{\partial^2 F^2}{\partial x \partial \varepsilon}(0,0) \neq 0, && \frac{\partial^2 F^2}{\partial x^2}(0,0) = 0, && \frac{\partial^3 F^2}{\partial x^3}(0,0) \neq 0. \end{aligned} \tag{17.83}$$

Since $\dfrac{\partial F^2}{\partial x}(0,0) = +1$ holds (because of $\dfrac{\partial F}{\partial x}(0,0) = -1$), the conditions for a *pitchfork bifurcation* are formulated for the mapping F^2.

For the differential equation (17.67) the properties of the mapping (17.82) imply that at $\alpha = 0$ a stable periodic orbit γ_α splits from γ with approximately a double period (*period doubling*), where γ loses its stability **(Fig. 17.30c)**.

■ The logistic mapping $\varphi_\alpha \colon [0,1] \to [0,1]$ is given for $0 < \alpha \leq 4$ by $\varphi_\alpha(x) = \alpha x(1-x)$, i.e., by the discrete dynamical system

$$x_{t+1} = \alpha x_t(1-x_t). \tag{17.84}$$

The mapping has the following bifurcation behavior (see [17.10]): For $0 < \alpha \leq 1$ system (17.84) has the equilibrium point 0 with domain of attraction $[0,1]$. For $1 < \alpha < 3$, (17.84) has the unstable equilibrium point 0 and the stable equilibrium point $1 - 1/\alpha$, where this last one has the domain of attraction $(0,1)$. For $\alpha_1 = 3$ the equilibrium point $1-1/\alpha$ is unstable and leads to a stable two-periodic orbit.

At the value $\alpha_2 = 1+\sqrt{6}$, the two-periodic orbit is also unstable and leads to a stable 2^2-periodic orbit. The period doubling continues, and stable 2^q-periodic orbits arise at $\alpha = \alpha_q$. Numerical evidence shows that $\alpha_q \to \alpha_\infty \approx 3.570\ldots$ as $q \to +\infty$.

For $\alpha = \alpha_\infty$, there is an attractor F (the *Feigenbaum attractor*), which has a structure similar to the Cantor set. There are points arbitrarily close to the attractor which are not iterated towards the attractor, but towards an unstable periodic orbit. The attractor F has dense orbits and the Hausdorff dimension is $d_H(F) \approx 0.538\ldots$. On the other hand, the dependence on initial conditions is not sensitive. In the domain $\alpha_\infty < \alpha < 4$, there exists a parameter set A with positive Lebesgue measure such that system (17.84) has a chaotic attractor of positive Lebesgue measure for $\alpha \in A$. The set A is interspersed with windows in which period doublings occur.

The bifurcation behavior of the logistic mapping can also be found in a class of *unimodal mappings*, i.e., of mappings of the interval I into itself, which has a single maximum in I. Although the param-

eter values α_i, for which period doubling occurs, are different from each other for different unimodal mappings, the rate of convergence by which these parameters tend to α_∞, is equal: $\alpha_k - \alpha_\infty \approx C\delta^{-k}$, where $\delta = 4.6692\ldots$ is the Feigenbaum constant (C depends on the concrete mapping). The Hausdorff dimension is the same for all attractors F at $\alpha = \alpha_\infty$: $d_H(F) \approx 0.538\ldots$.

4. Creation of a Torus

Consider system (17.67) with $n \geq 3, r \geq 6$ and $l = 1$. Suppose that for all ε close to 0 system (17.67) has a periodic orbit γ_ε. Let the multipliers of γ_0 be $\rho_{1,2} = e^{\pm i\Psi}$ with $\Psi \notin \left\{0, \frac{\pi}{2}, \frac{2\pi}{3}, \pi\right\}$, ρ_j $(j = 3, \ldots, n-1)$ with $|\rho_j| \neq 1$ and $\rho_n = 1$.
According to the center manifold theorem, in this case there exists a two-dimensional reduced C^6-mapping

$$x \longmapsto F(x, \varepsilon) \tag{17.85}$$

with $F(0, \varepsilon) = 0$ for ε close to 0.
If the Jacobian matrix $D_x F(0, \varepsilon)$ has the conjugate complex eigenvalues $\rho(\varepsilon)$ and $\overline{\rho}(\varepsilon)$ with $|\rho(0)| = 1$ for all ε near 0, if $d := \frac{d}{d\varepsilon}|\rho(\varepsilon)|_{|\varepsilon=0} > 0$ holds and $\rho(0)$ is not a q-th root of 1 for $q = 1, 2, 3, 4$, then (17.85) can be transformed by a smooth ε dependent coordinate transformation into the form $x \mapsto \tilde{F}(x, \varepsilon) = \tilde{F}_o(x, \varepsilon) + O(\|x\|^5)$ (O Landau symbol), where $\tilde{F}_o$ is given in polar coordinates by

$$\begin{pmatrix} r \\ \vartheta \end{pmatrix} \longmapsto \begin{pmatrix} |\rho(\varepsilon)|r + a(\varepsilon)r^3 \\ \vartheta + \omega(\varepsilon) + b(\varepsilon)r^2 \end{pmatrix}. \tag{17.86}$$

Here α, ω and b are differentiable functions. Suppose $a(0) < 0$ holds. Then, the equilibrium point $r = 0$ of (17.86) is asymptotically stable for all $\varepsilon < 0$ and unstable for $\varepsilon > 0$. Furthermore, for $\varepsilon > 0$ there exists the circle $r = \sqrt{-\frac{d\varepsilon}{a(0)}}$, which is invariant under the mapping (17.86) and asymptotically stable **(Fig. 17.31a)**.

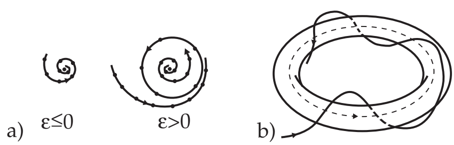


Figure 17.31

The Neimark-Sacker Theorem (see [17.10], [17.1]) states that the bifurcation behavior of (17.86) is similar to that of $\tilde{F}$ (*supercritical Hopf bifurcation for mappings*).

■ In mapping (17.85), given by

$$\begin{pmatrix} x \\ y \end{pmatrix} \longmapsto \frac{1}{\sqrt{2}} \begin{pmatrix} (1+\varepsilon)x + y + x^2 - 2y^2 \\ -x + (1+\varepsilon)y + x^2 - x^3 \end{pmatrix},$$

there is a supercritical Hopf bifurcation at $\varepsilon = 0$.

With respect to the differential equation (17.67), the existence of a closed invariant curve of mapping (17.85) means that the periodic orbit γ_0 is unstable for $a(0) < 0$ and for $\varepsilon > 0$ a torus arises which is invariant with respect to (17.67) **(Fig. 17.31b)**.

17.3.1.3 Global Bifurcation

Besides the periodic creation orbit which arises if a separatrix loop disappears, (17.67) can have further global bifurcations. Two of them are shown in [17.12] by examples.

1. Emergence of a Periodic Orbit due to the Disappearance of a Saddle-Node

■ The parameter-dependent system

$$\dot{x} = x(1 - x^2 - y^2) + y(1 + x + \alpha),\ \dot{y} = -x(1 + x + \alpha) + y(1 - x^2 - y^2)$$

has in polar coordinates $x = r\cos\vartheta,\ y = r\sin\vartheta$ the following form:

$$\dot{r} = r(1 - r^2),\ \dot{\vartheta} = -(1 + \alpha + r\cos\vartheta). \tag{17.87}$$

Obviously, the circle $r = 1$ is invariant under (17.87) for an arbitrary parameter α, and all orbits (except the equilibrium point $(0, 0)$) tend to this circle for $t \to +\infty$. For $\alpha < 0$ there are a saddle and a stable node on the circle, which fuse into a compound saddle-node type equilibrium point at $\alpha = 0$. For $\alpha > 0$, there is no equilibrium point on the circle, which is a periodic orbit **(Fig. 17.32)**.

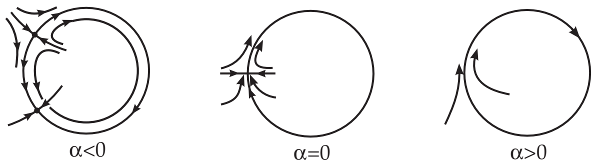


Figure 17.32

2. Disappearance of a Saddle-Saddle Separatrix in the Plane

■ Consider the parameter-dependent planar differential equation

$$\dot{x} = \alpha + 2xy,\ \dot{y} = 1 + x^2 - y^2. \tag{17.88}$$

For $\alpha = 0$, equation (17.88) has the two saddles $(0, 1)$ and $(0, -1)$ and the y-axis as invariant sets. The heteroclinic orbit is part of this invariant set. For small $|\alpha| \neq 0$, the saddle-points are retained while the heteroclinic orbit disappears **(Fig. 17.33)**.

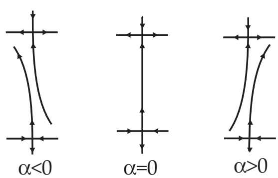


Figure 17.33

17.3.2 Transitions to Chaos

Often a strange attractor does not arise suddenly but as the result of a sequence of bifurcations, from which the typical ones are represented in Section 17.3.1. The most important ways to create strange attractors or strange invariant sets are described in the following.

17.3.2.1 Cascade of Period Doublings

Analogously to the logistic equation (17.84), a cascade of period doublings can also occur in time-continuous systems in the following way. Suppose system (17.67) has the stable periodic orbit $\gamma_\varepsilon^{(1)}$ for $\varepsilon < \varepsilon_1$. For $\varepsilon = \varepsilon_1$ a period doubling occurs near $\gamma_{\varepsilon_1}^{(1)}$, at which the periodic orbit $\gamma_\varepsilon^{(1)}$ loses its stability for $\varepsilon > \varepsilon_1$. A periodic orbit $\gamma_{\varepsilon_1}^{(2)}$ with approximately double period splits from it. At $\varepsilon = \varepsilon_2$, there is a

new period doubling, where $\gamma_{\varepsilon_2}^{(2)}$ loses its stability and a stable orbit $\gamma_{\varepsilon_2}^{(4)}$ with an approximately double period arises. For important classes of systems (17.67) this procedure of period doubling continues, so a sequence of parameter values $\{\varepsilon_j\}$ arises.

Numerical calculations for certain differential equations (17.67), e.g., for hydrodynamical differential equations such as the Lorenz system, show the existence of the limit

$$\lim_{j\to+\infty} \frac{\varepsilon_{j+1}-\varepsilon_j}{\varepsilon_{j+2}-\varepsilon_{j+1}} = \delta, \text{ where } \delta \tag{17.89}$$

is again the Feigenbaum constant.

For $\varepsilon_* = \lim\limits_{j\to\infty} \varepsilon_j$, the cycle with infinite period loses its stability, and a strange attractor appears.

The geometric background for this strange attractor in (17.67) by a cascade of period doubling is shown in **Fig. 17.34a**. The Poincaré section shows approximately a baker mapping, which suggests that a Cantor set-like structure is created.

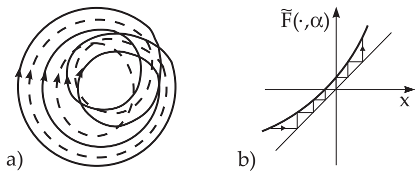


Figure 17.34

17.3.2.2 Intermittency

Consider a stable periodic orbit of (17.67), which loses its stability for $\varepsilon = 0$ when exactly one of the multipliers, for $\varepsilon < 0$ inside the unit circle takes the value $+1$. According to the center manifold theorem, the corresponding saddle-node bifurcation of the Poincaré mapping can be described by a one-dimensional mapping in the normal form $x \mapsto \tilde{F}(x,\alpha) = \alpha + x + x^2 + \cdots$. Here α is a parameter depending on ε, i.e., $\alpha = \alpha(\varepsilon)$ with $\alpha(0) = 0$. The graph of $\tilde{F}(\cdot,\alpha)$ for positive α is represented in **Fig. 17.34b**.

As can be seen in **Fig. 17.34b**, the iterates of $\tilde{F}(\cdot,\alpha)$ stay for a relatively long time in the tunnel zone for $\alpha \gtrsim 0$. For equation (17.67), this means that the corresponding orbits stay relatively long in the neighborhood of the original periodic orbit. During this time, the behavior of (17.68) is approximately periodic (*laminar phase*). After getting through the tunnel zone the considered orbit escapes, which results in irregular motion (*turbulent phase*). After a certain time the orbit is recovered and a new laminar phase begins. A strange attractor arises in this situation if the periodic orbit vanishes and its stability goes over to the chaotic set. The saddle-node bifurcation is only one of the typical local bifurcations playing a role in the intermittence scenario. Two further ones are period doubling and the creation of a torus.

17.3.2.3 Global Homoclinic Bifurcations

1. Smale's Theorem

Let the invariant manifolds of the Poincaré mapping of the differential equation (17.67) in $\mathbb{R}^3$ near the periodic orbit γ be as in **Fig. 17.11b**, p. 873. The transversal homoclinic points $P^j(x_0)$ correspond to a homoclinic orbit of (17.67) to γ. The existence of such a homoclinic orbit in (17.67) leads to a sensitive dependence on initial conditions. In connection with the considered Poincaré mapping, horseshoe mappings, introduced by Smale, can be constructed. This leads to the following statements:

a) In every neighborhood of a transversal homoclinic point of the Poincaré mapping (17.80) there exists a periodic point of this mapping (*Smale's theorem*). Hence, in every neighborhood of a transversal

homoclinic point there exists an invariant set of $P^m (m \in \mathbb{N})$, Λ, which is of Cantor type. The restriction of P^m to Λ is topologically conjugate to a Bernoulli shift, i.e., to a mixing system.

b) The invariant set of the differential equation (17.67) close to the homoclinic orbit is like a product of a Cantor set with the unit circle. If this invariant set is attracting, then it represents a strange attractor of (17.67).

2. Shilnikov's Theorem

Consider the differential equation (17.67) in $\mathbb{R}^3$ with scalar parameter ε. Suppose that the system (17.67) has a saddle-node type hyperbolic steady state 0 at $\varepsilon = 0$, which exists so long as $|\varepsilon|$ remains small. Suppose, that the Jacobian matrix $D_x f(0,0)$ has the eigenvalue $\lambda_3 > 0$ and a pair of conjugate complex eigenvalues $\lambda_{1,2} = a \pm \mathrm{i}\,\omega$ with $a < 0$. Suppose, additionally, that (17.67) has a separatrix loop γ_0 for $\varepsilon = 0$, i.e., a homoclinic orbit which tends to 0 for $t \to -\infty$ and $t \to +\infty$ **(Fig. 17.35a)**.

Then, in a neighborhood of a separatrix loop (17.67) has the following phase portrait :

a) Let $\lambda_3 + a < 0$. If the separatrix loop breaks at $\varepsilon \neq 0$ according to the variant denoted by A in **(Fig. 17.35a)**, then there is exactly one periodic orbit of (17.67) for $\varepsilon = 0$. If the separatrix loop breaks at $\varepsilon \neq 0$ according to the variant denoted by B in **(Fig. 17.35a)**, then there is no periodic orbit.

b) Let $\lambda_3 + a > 0$. Then there exist countably many saddle-type periodic orbits at $\varepsilon = 0$ (respectively, for small $|\varepsilon|$) close to the separatrix loop γ_0 (respectively, close to the destructed loop γ_0). The Poincaré mapping with respect to a transversal to the γ_0 plane generates a countable set of horseshoe mappings at $\varepsilon = 0$, from which there remain finitely many for small $|\varepsilon| \neq 0$.

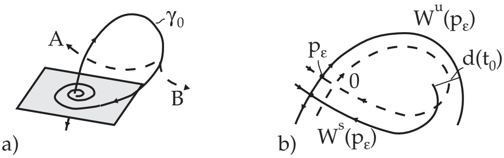


Figure 17.35

3. Melnikov's Method

Consider the planar differential equation

$$\dot{x} = f(x) + \varepsilon g(t, x), \tag{17.90}$$

where ε is a small parameter. For $\varepsilon = 0$, let (17.90) be a Hamiltonian system (see 17.1.4.3, **2.**, p. 875), i.e., for $f = (f_1, f_2)$ $f_1 = \dfrac{\partial H}{\partial x_2}$ and $f_2 = -\dfrac{\partial H}{\partial x_1}$ hold, where $H\colon U \subset \mathbb{R}^2 \to \mathbb{R}$ is supposed to be a C^3-function. Suppose the time-dependent vector field $g\colon \mathbb{R} \times U \to \mathbb{R}^2$ is twice continuously differentiable, and T-periodic with respect to the first argument. Furthermore, let f and g be bounded on bounded sets. Suppose that for $\varepsilon = 0$ there exists a homoclinic orbit with respect to the saddle point 0, and the Poincaré section $\sum_{t_0}$ of (17.90) in the phase space $\{(x_1, x_2, t)\}$ for $t = t_0$ looks as in **Fig. 17.35b**. The Poincaré mapping $P_{\varepsilon,t_0}: \sum_{t_0} \to \sum_{t_0}$, for small $|\varepsilon|$, has a saddle point p_ε close to $x = 0$ with the invariant manifolds $W^s(p_\varepsilon)$ and $W^u(p_\varepsilon)$. If the homoclinic orbit of the unperturbed system is given by $\varphi(t - t_0)$, then the distance between the manifold $W^s(p_\varepsilon)$ and $W^u(p_\varepsilon)$, measured along the line passing through $\varphi(0)$ and perpendicular to $f(\varphi(0))$, can be calculated by the formula

$$d(t_0) = \varepsilon \frac{M(t_0)}{\| f(\varphi(0)) \|} + O(\varepsilon^2). \tag{17.91a}$$

Here, $M(\cdot)$ is the *Melnikov function* which is defined by

$$M(t_0) = \int_{-\infty}^{+\infty} f(\varphi(t - t_0)) \wedge g(t, \varphi(t - t_0))\, dt. \tag{17.91b}$$

(For $a = (a_1, a_2)$ and $b = (b_1, b_2)$, $\wedge$ means $a \wedge b = a_1 b_2 - a_2 b_1$.) If the Melnikov function M has a simple root at t_0, i.e., $M(t_0) = 0$ and $M'(t_0) \neq 0$ hold, then the manifolds $W^s(p_\varepsilon)$ and $W^u(p_\varepsilon)$ intersect each other transversally for sufficiently small $\varepsilon > 0$. If M has no root, then $W^s(p_\varepsilon) \cap W^u(p_\varepsilon) = \emptyset$, i.e., there is no homoclinic point.

Remark: Suppose the unperturbed system (17.90) has a heteroclinic orbit given by $\varphi(t - t_0)$, running from a saddle point 0_1 in a saddle 0_2. Let p_ε^1 and p_ε^2 be the saddle points of the Poincaré mapping P_{ε,t_0} for small $|\varepsilon|$. If M, calculated as above, has a simple root at t_0, then $W^s(p_\varepsilon^1)$ and $W^u(p_\varepsilon^2)$ intersect each other transversally for small $\varepsilon > 0$.

■ Consider the periodically perturbed pendulum equation $\ddot{x} + \sin x = \varepsilon \sin \omega t$, i.e., the system $\dot{x} = y,\ \dot{y} = -\sin x + \varepsilon \sin \omega t$, in which ε is a small parameter and ω is a further parameter. The unperturbed system $\dot{x} = y,\ \dot{y} = -\sin x$ is a Hamiltonian system with $H(x,y) = \frac{1}{2}y^2 - \cos x$. It has (among others) a pair of heteroclinic orbits through $(-\pi, 0)$ and $(\pi, 0)$ (in the cylindrical phase space $S^1 \times \mathbb{R}$ these are homoclinic orbits) given by $\varphi^\pm(t) = \left(\pm 2\arctan(\sinh t), \pm 2\frac{1}{\cosh t}\right)$ $(t \in \mathbb{R})$. The direct calculation of the Melnikov function yields $M(t_0) = \mp \dfrac{2\pi \sin \omega t_0}{\cosh(\pi\omega/2)}$. Since M has a simple root at $t_0 = 0$, the Poincaré mapping of the perturbed system has transversal homoclinic points for small $\varepsilon > 0$.

17.3.2.4 Destruction of a Torus

1. From Torus to Chaos

1. Hopf-Landau Model of Turbulence The problem of transition from regular (laminar) behavior to irregular (turbulent) behavior is especially interesting for systems with distributed parameters, which are described, e.g., by partial differential equations. From this viewpoint, *chaos* can be interpreted as behavior irregular in time but ordered in space.

On the other hand, *turbulence* is the behavior of the system, that is irregular in time and in space. The Hopf-Landau model explains the arising of turbulence by an infinite cascade of Hopf bifurcations: For $\varepsilon = \varepsilon_1$ a steady state bifurcates in a limit cycle, which becomes unstable for $\varepsilon_2 > \varepsilon_1$ and leads to a torus T^2. At the k-th bifurcation of this type a k-dimensional torus arises, generated by non-closed orbits which wind on it. The Hopf-Landau model does not lead in general to an attractor which is characterized by sensitive dependence on the initial conditions and mixing.

2. Ruelle-Takens-Newhouse Scenario Suppose that in system (17.67) $n \geq 4$ and $l = 1$ hold. Suppose also that changing the parameter ε, the bifurcation sequence "equilibrium point $\to$ periodic orbit $\to$ torus $T^2 \to$ torus T^3" is achieved by three consecutive Hopf bifurcations.
Let the quasiperiodic flow on T^3 be structurally unstable. Then, certain small perturbation of (17.67) can lead to the destruction of T^3 and to the creation of a structurally stable strange attractor.

3. Theorem of Afraimovich and Shilnikov on the Loss of Smoothness and the Destruction of the Torus T^2 Let the sufficiently smooth system (17.67) be given with $n \geq 3$ and $l = 2$. Suppose that for the parameter value ε_0, the system (17.67) has an attracting smooth torus $T^2(\varepsilon_0)$ spanned by a stable periodic orbit γ_s, a saddle-type periodic orbit γ_u and its unstable manifold $W^u(\gamma_u)$ (*resonance torus*).
The invariant manifolds of the equilibrium points of the Poincaré mapping computed with respect to a surface transversal to the torus in the longitudinal direction, are represented in **Fig. 17.36a**. The multiplier ρ of the orbit γ_s, which is the nearest to the unit circle, is assumed to be real and simple. Furthermore, let $\varepsilon(\cdot) : [0,1] \to V$ be an arbitrary continuous curve in parameter space, for which $\varepsilon(0) = \varepsilon_0$ and for which system (17.67) has no invariant resonance torus for $\varepsilon = \varepsilon(1)$. Then the following statements are true:

a) There exists a value $s_* \in (0,1)$ for which $T^2(\varepsilon(s_*))$ loses its smoothness. Here, either the multiplier $\rho(s_*)$ is complex or the unstable manifold $W^u(\gamma_u)$ loses its smoothness close to γ_s.

b) There exists a further parameter value $s_{**} \in (s_*, 1)$ such that system (17.67) has no resonance torus for $s \in (s_{**}, 1]$. The torus is destroyed in the following way:

$\boldsymbol{\alpha}$) The periodic orbit γ_s loses its stability for $\varepsilon = \varepsilon(s_{**})$. A local bifurcation arises as period doubling or the formation of a torus.

$\boldsymbol{\beta}$) The periodic orbits γ_u and γ_s coincide for $\varepsilon = \varepsilon(s_{**})$ (saddle-node bifurcation) and so they vanish.

$\boldsymbol{\gamma}$) The stable and unstable manifolds of γ_u intersect each other non-transversally for $\varepsilon = \varepsilon(s_{**})$ (see the *bifurcation diagram* in **Fig. 17.36c**). The points of the beak-shaped curve S_1 correspond to the fused γ_s and γ_u (saddle-node bifurcation). The tip C_1 of the beak-shaped curve is on a curve S_0, which corresponds to a splitting of the torus.

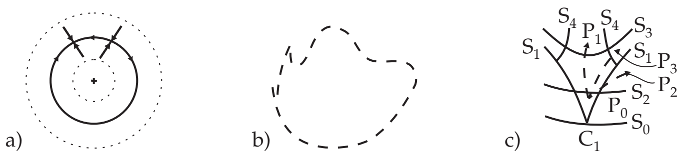


Figure 17.36

The parameter points where the smoothness is lost, are on the curve S_2, while the points on S_3 characterize the dissolving of a T^2 torus. The parameter points for which the stable and unstable manifolds of γ_u intersect each other non-transversally, are on the curve S_4. Let P_0 be an arbitrary point in the beaked shaped tip of the beak such that for this parameter value a resonance torus T^2 arises. The transition from P_0 to P_1 corresponds to the case $\boldsymbol{\alpha}$) of the theorem. If the multiplier ρ becomes -1 on S_2, then there is a period doubling. A cascade of further period doublings can lead to a strange attractor. If a pair of complex conjugate multipliers $\rho_{1,2}$ arises on the unit circle passing through S_2, then it can result in the splitting of a further torus, for which the Afraimovich-Shilnikov theorem can be used again.
The transition from P_0 to P_2 represents the case $\boldsymbol{\beta}$) of the theorem: The torus loses its smoothness, and on passing through on S_1, there is a saddle-node bifurcation. The torus is destroyed, and a transition to chaos through intermittence can happen. The transition from P_0 to P_3, finally, corresponds to the case $\boldsymbol{\gamma}$): After the loss of smoothness, a non-robust homoclinic curve forms on passing through on S_4. The stable cycle γ_s remains and a hyperbolic set arises which is not attracting for the present. If γ_s vanishes, then a strange attractor arises from this set.

2. Mappings on the Unit Circle and Rotation Number

1. Equivalent and Lifted Mappings The properties of the invariant curves of the Poincaré mapping play an important role in the loss of smoothness and destruction of a torus. If the Poincaré mapping is represented in polar coordinates, then, under certain assumptions, one gets decoupled mappings of the angular variables as informative auxiliary mappings on the unit circle. These are invertible in the case of smooth invariant curves **(Fig. 17.36a)** and in the case of non-smooth curves **(Fig. 17.36b)** they are not invertible. A mapping $F: \mathbb{R} \to \mathbb{R}$ with $F(\Theta + 1) = F(\Theta) + 1$ for all $\Theta \in \mathbb{R}$, which generates the dynamical system

$$\Theta_{n+1} = F(\Theta_n), \tag{17.92}$$

is called *equivariant*. For every such mapping, an associated mapping of the unit circle $f: S^1 \to S^1$ can be assigned where $S^1 = \mathbb{R} \setminus \mathbb{Z} = \{\Theta \mod 1, \Theta \in \mathbb{R}\}$. Here $f(x) := F(\Theta)$ if the relation $x = [\Theta]$ holds for the equivalence class $[\Theta]$. F is called a *lifted mapping of f*. Obviously, this construction is not unique. In contrast to (17.92)

$$x_{t+1} = f(x_t) \tag{17.93}$$

is a dynamical system on S^1.

■ For two parameters ω and K let the mapping $\tilde{F}(\cdot\,; \omega,\ K)$ be defined by $\tilde{F}(\sigma; \omega, K) = \sigma + \omega - K \sin \sigma$

for all $\tau \in \mathbb{R}$. The corresponding dynamical system

$$\sigma_{n+1} = \sigma_n + \omega - K \sin \sigma_n \tag{17.94}$$

can be transformed by the transformation $\sigma_n = 2\pi\Theta_n$ into the system

$$\Theta_{n+1} = \Theta_n + \Omega - \frac{K}{2\pi} \sin 2\pi\Theta_n \tag{17.95}$$

where $\Omega = \frac{\omega}{2\pi}$. With $F(\Theta;\Omega,K) = \Theta+\Omega-\frac{K}{2\pi}\sin 2\pi\Theta$ an equivariant mapping arises, which generates the *canonical form of the circle mapping.*

2. Rotation Number The orbit $\gamma(\Theta) = \{F^n(\Theta)\}$ of (17.92) is a *q-periodic orbit* of (17.93) in S^1 if and only if it is a $\frac{p}{q}$ cycle of (17.92), i.e., if there exists an integer p such that $\Theta_{n+q} = \Theta_n + p, (n \in \mathbb{Z})$ holds. The mapping $f\colon S^1 \to S^1$ is called *orientation preserving* if there exists a corresponding lifted mapping F, which is monotone increasing. If F from (17.92) is a monotone increasing homeomorphism, then there exists the limit $\lim\limits_{|n|\to\infty} \frac{F^n(x)}{n}$ for every $x \in \mathbb{R}$, and this limit does not depend on x. Hence, the expression $\rho(F) := \lim\limits_{|n|\to\infty} \frac{F^n(x)}{n}$ can be defined. If $f\colon S^1 \to S^1$ is a homeomorphism and F and $\tilde{F}$ are two lifted mappings of f, then $\rho(F) = \rho(\tilde{F}) + k$ holds, where k is an integer. Based on this last property, the *rotation number* $\rho(f)$ of an orientation-preserving homeomorphism $f\colon S^1 \to S^1$ can be defined as $\rho(f) = \rho(F) \bmod 1$, where F is an arbitrary lifted mapping of f.

If $f\colon S^1 \to S^1$ in (17.93) is an orientation-preserving homeomorphism, then the rotation number has the following properties (see [17.4]):

a) If (17.93) has a q-periodic orbit, then there exists an integer p such that $\rho(f) = \frac{p}{q}$ holds.

b) If $\rho(f) = 0$, then (17.93) has an equilibrium point.

c) If $\rho(f) = \frac{p}{q}$, where $p \neq 0$ is an integer and q is a natural number (p and q are co-primes), then (17.93) has a q-periodic orbit.

d) $\rho(f)$ is irrational if and only if (17.93) has neither a periodic orbit nor an equilibrium point.

Theorem of Denjoy: If $f\colon S^1 \to S^1$ is an orientation-preserving C^2-diffeomorphism and the rotation number $\alpha = \rho(f)$ is irrational, then f is topologically conjugate to a pure rotation whose lifted mapping is $F(x) = x + \alpha$.

3. Differential Equations on the Torus T^2

Let

$$\dot{\Theta}_1 = f_1(\Theta_1, \Theta_2), \quad \dot{\Theta}_2 = f_2(\Theta_1, \Theta_2) \tag{17.96}$$

be a planar differential equation, where f_1 and f_2 are differentiable and 1-periodic functions in both arguments. In this case (17.96) defines a flow, which can also be interpreted as a flow on the torus $T^2 = S^1 \times S^1$ with respect to Θ_1 and Θ_2. If $f_1(\Theta_1, \Theta_2) > 0$ for all (Θ_1, Θ_2), then (17.96) has no equilibrium points and it is equivalent to the scalar first-order differential equation

$$\frac{d\Theta_2}{d\Theta_1} = \frac{f_2(\Theta_1, \Theta_2)}{f_1(\Theta_1, \Theta_2)}. \tag{17.97}$$

With the relations $\Theta_1 = t$, $\Theta_2 = x$ and $f = \frac{f_2}{f_1}$, (17.97) can be written as a non-autonomous differential equation

$$\dot{x} = f(t, x) \tag{17.98}$$

whose right-hand side has a period of 1 with respect to t and x.
Let $\varphi(\cdot, x_0)$ be the solution of (17.98) with initial state x_0 at time $t = 0$. So, a mapping $\varphi^1(\cdot) = \varphi(1, \cdot)$ can be defined for (17.98), which can be considered as the lifted mapping of a mapping $f\colon S^1 \to S^1$.

■ Let $\omega_1, \omega_2 \in \mathbb{R}$ be constants and $\dot{\Theta}_1 = \omega_1$, $\dot{\Theta}_2 = \omega_2$ a differential equation on the torus, which is equivalent to the scalar differential equation $\dot{x} = \dfrac{\omega_2}{\omega_1}$ for $\omega_1 \neq 0$. Thus, $\varphi(t, x_0) = \dfrac{\omega_2}{\omega_1}t + x_0$ and $\varphi^1(x) = \dfrac{\omega_2}{\omega_1} + x$.

4. Canonical Form of a Circle Mapping

1. Canonical Form The mapping F from (17.95) is an orientation-preserving diffeomorphism for $0 \le K < 1$, because $\dfrac{\partial F}{\partial \vartheta} = 1 - K\cos 2\pi\vartheta > 0$ holds. For $K = 1$, F is no-longer a diffeomorphism, but it is still a homeomorphism, while for $K > 1$, the mapping is not invertible, and hence no-longer a homeomorphism. In the parameter domain $0 \le K \le 1$, the rotation number $\rho(\Omega, K) := \rho(F(\cdot, \Omega, K))$ is defined for $F(\cdot, \Omega, K)$. Let $K \in (0, 1)$ be fixed. Then $\rho(\cdot, K)$ has the following properties on [0,1]:

a) The function $\rho(\cdot, K)$ is not decreasing, it is continuous, but it is not differentiable.

b) For every rational number $\dfrac{p}{q} \in [0, 1)$ there exists an interval $I_{p/q}$, whose interior is not empty and for which $\rho(\Omega, K) = \dfrac{p}{q}$ holds for all $\Omega \in I_{p/q}$.

c) For every irrational number $\alpha \in (0, 1)$ there exists exactly one Ω with $\rho(\Omega, K) = \alpha$.

2. Devil's Staircase and Arnold Tongues For every $K \in (0, 1)$, $\rho(\cdot, K)$ is a Cantor function. The graph of $\rho(\cdot, K)$, which is represented in **Fig. 17.37b**, is called the *devil's staircase*. The bifurcation diagram of (17.95) is represented in **Fig. 17.37a**. At every rational number on the Ω-axis, a beak-shaped region (*Arnold tongue*) with a non-empty interior starts, where the rotation number is constant and equal to the rational number.

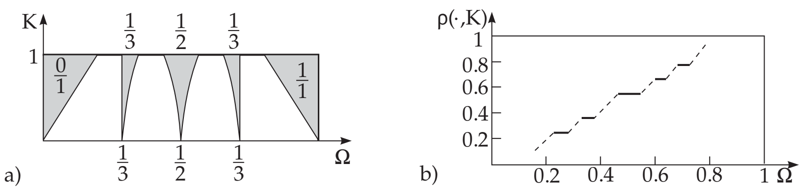


Figure 17.37

The reason for the formation of the tongue is a synchronization of the frequencies (*frequency locking*).

a) For $0 \le K < 1$, these regions are not overlapping. At every irrational number of the Ω-axis, a continuous curve starts which always reaches the line $K = 1$. In the first Arnold tongue with $\rho = 0$, the dynamical system (17.95) has equilibrium points. If K is fixed and Ω increases, then two of these equilibrium points fuse on the boundary of the first Arnold tongue and vanish at the same time. As a result of such a saddle-node bifurcation, a dense orbit arises on S^1. Similar phenomena can be observed when leaving other Arnold tongues.

b) For $K > 1$ the theory of the rotation numbers is no-longer applicable. The dynamics become more complicated, and the transition to chaos takes place. Here, similarly to the case of Feigenbaum constants, further constants arise, which are equal for certain classes of mappings to which also the standard circle mapping belongs. One of them is described below.

3. Golden Mean, Fibonacci Numbers The irrational number $\dfrac{\sqrt{5}-1}{2}$ is called the *golden mean* and it has a simple continued fraction representation $\dfrac{\sqrt{5}-1}{2} = \dfrac{1}{1+\frac{1}{1+\frac{1}{1+\cdots}}} = [1;1,1,\ldots]$ (see 1.1.1.4, **3.**, p. 4). By successive evaluation of the continued fraction a sequence $\{r_n\}$ of rational numbers is got, which converges to $\dfrac{\sqrt{5}-1}{2}$. The numbers r_n can be represented in the form $r_n = \dfrac{F_n}{F_{n+1}}$, where F_n are Fibonacci numbers (see 5.4.1.5, p. 375), which are determined by the iteration $F_{n+1} = F_n + F_{n-1}$ $(n = 1, 2, \cdots)$ with initial values $F_0 = 0$ and $F_1 = 1$. Now, let Ω_∞ be the parameter value of (17.95), for which $\rho(\Omega_\infty, 1) = \dfrac{\sqrt{5}-1}{2}$ and let Ω_n be the closest value to Ω_∞, for which $\rho(\Omega_n, 1) = r_n$ holds. Numerical calculation gives the limit $\lim\limits_{n\to\infty} \dfrac{\Omega_n - \Omega_{n-1}}{\Omega_{n+1} - \Omega_n} = -2.8336\ldots$.

18 Optimization

18.1 Linear Programming

18.1.1 Formulation of the Problem and Geometrical Representation

18.1.1.1 The Form of a Linear Programming Problem

1. The Subject

of *linear programming* is the minimization or maximization of a *linear objective function* (**OF**) of finitely many variables subject to a finite number of *constraints* (**CT**), which are given as linear equations or inequalities.
Many practical problems can be directly formulated as a linear programming problem, or they can be modeled approximately by a linear programming problem.

2. General Form

A linear programming problem has the following general form:

$$\textbf{OF:}\quad f(\underline{\mathbf{x}}) = c_1x_1 + \cdots + c_rx_r + c_{r+1}x_{r+1} + \cdots + c_nx_n = \max! \tag{18.1a}$$

$$\textbf{CT:}\quad \left.\begin{array}{ccccccccc} a_{1,1}x_1 & +\cdots+ & a_{1,r}x_r & + & a_{1,r+1}x_{r+1} & +\cdots+ & a_{1,n}x_n & \le & b_1, \\ \vdots & & \vdots & & \vdots & & \vdots & & \vdots \\ a_{s,1}x_1 & +\cdots+ & a_{s,r}x_r & + & a_{s,r+1}x_{r+1} & +\cdots+ & a_{s,n}x_n & \le & b_s, \\ a_{s+1,1}x_1 & +\cdots+ & a_{s+1,r}x_r & + & a_{s+1,r+1}x_{r+1} & +\cdots+ & a_{s+1,n}x_n & = & b_{s+1}, \\ \vdots & & \vdots & & \vdots & & \vdots & & \vdots \\ a_{m,1}x_1 & +\cdots+ & a_{m,r}x_r & + & a_{m,r+1}x_{r+1} & +\cdots+ & a_{m,n}x_n & = & b_m, \\ & & & & & & & & \\ \multicolumn{9}{l}{x_1 \ge 0, \ldots, x_r \ge 0; \qquad x_{r+1}, \ldots, x_n \quad \text{free.}} \end{array}\right\} \tag{18.1b}$$

In a more compact vector notation this problem becomes:

$$\textbf{OF}:\quad f(\underline{\mathbf{x}}) = \underline{\mathbf{c}}^{1\mathrm{T}}\underline{\mathbf{x}}^1 + \underline{\mathbf{c}}^{2\mathrm{T}}\underline{\mathbf{x}}^2 = \max! \tag{18.2a}$$

$$\textbf{CT}:\quad \left.\begin{array}{l} \mathbf{A}_{11}\underline{\mathbf{x}}^1 + \mathbf{A}_{12}\underline{\mathbf{x}}^2 \le \underline{\mathbf{b}}^1, \\ \mathbf{A}_{21}\underline{\mathbf{x}}^1 + \mathbf{A}_{22}\underline{\mathbf{x}}^2 = \underline{\mathbf{b}}^2, \\ \underline{\mathbf{x}}^1 \ge 0,\ \underline{\mathbf{x}}^2 \ \text{free.} \end{array}\right\} \tag{18.2b}$$

Here, the following notations are used:

$$\underline{\mathbf{c}}^1 = \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_r \end{pmatrix},\ \underline{\mathbf{c}}^2 = \begin{pmatrix} c_{r+1} \\ c_{r+2} \\ \vdots \\ c_n \end{pmatrix}, \qquad \underline{\mathbf{x}}^1 = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_r \end{pmatrix},\ \underline{\mathbf{x}}^2 = \begin{pmatrix} x_{r+1} \\ x_{r+2} \\ \vdots \\ x_n \end{pmatrix}, \tag{18.2c}$$

$$\mathbf{A}_{11} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1,r} \\ a_{21} & a_{22} & \cdots & a_{2,r} \\ \vdots & & & \\ a_{s,1} & a_{s,2} & \cdots & a_{s,r} \end{pmatrix}, \qquad \mathbf{A}_{12} = \begin{pmatrix} a_{1,r+1} & a_{1,r+2} & \cdots & a_{1,n} \\ a_{2,r+1} & a_{2,r+2} & \cdots & a_{2,n} \\ \vdots & & & \\ a_{s,r+1} & a_{s,r+2} & \cdots & a_{s,n} \end{pmatrix}, \tag{18.2d}$$

$$\mathbf{A}_{21} = \begin{pmatrix} a_{s+1,1} & a_{s+1,2} & \cdots & a_{s+1,r} \\ a_{s+2,1} & a_{s+2,2} & \cdots & a_{s+2,r} \\ \vdots & & & \\ a_{m,1} & a_{m,2} & \cdots & a_{m,r} \end{pmatrix}, \qquad \mathbf{A}_{22} = \begin{pmatrix} a_{s+1,r+1} & a_{s+1,r+2} & \cdots & a_{s+1,n} \\ a_{s+2,r+1} & a_{s+2,r+2} & \cdots & a_{s+2,n} \\ \vdots & & & \\ a_{m,r+1} & a_{m,r+2} & \cdots & a_{m,n} \end{pmatrix}. \tag{18.2e}$$

3. Constraints

with the inequality sign " $\ge$ " will have the above form if they are multiplied by (-1).

I.N. Bronshtein et al., *Handbook of Mathematics*,
DOI 10.1007/978-3-662-46221-8_18

4. Minimum Problem

A minimum problem $f(\underline{\mathbf{x}}) = \min!$ becomes an equivalent maximum problem by multiplying the objective function by (-1)

$$-f(\underline{\mathbf{x}}) = \max! \tag{18.3}$$

5. Integer Programming

Sometimes certain variables are restricted to be only integers. This discrete problem is not discussed here.

6. Formulation with only Non-Negative Variables and Slack Variables

In applying certain solution methods, only non-negative variables are considered, and constraints (18.1b), (18.2b) given in equality form.

Every free variable x_k must be decomposed into the difference of two non-negative variables $x_k = x_k^1 - x_k^2$. The inequalities become equalities by adding non-negative variables; they are called *slack variables*. That is, the problem is considered in the form as given in (18.4a,b), where n is the increased number of variables. In vector form:

$$\textbf{OF}: f(\underline{\mathbf{x}}) = c_1x_1 + \cdots + c_nx_n = \max! \tag{18.4a}$$

$$\textbf{CT}: \left.\begin{array}{rcl} a_{1,1}x_1 + \cdots + a_{1,n}x_n &=& b_1, \\ \vdots \qquad\qquad \vdots && \vdots \\ a_{m,1}x_1 + \cdots + a_{m,n}x_n &=& b_m, \\ \\ x_1 \geq 0, \ldots, x_n \geq 0. && \end{array}\right\} \tag{18.4b}$$

OF: $\quad f(\underline{\mathbf{x}}) = \underline{\mathbf{c}}^{\mathrm{T}}\underline{\mathbf{x}} = \max!$ (18.5a) **CT:** $\quad \mathbf{A}\underline{\mathbf{x}} = \underline{\mathbf{b}}\,, \quad \underline{\mathbf{x}} \geq \underline{\mathbf{0}}\,.$ (18.5b)

The relation $m \leq n$ can be supposed, otherwise the system of equations contains linearly dependent or contradictory equations.

7. Feasible Set

The set of all vectors $\underline{\mathbf{x}}$ satisfying constraints (18.2b) is called the *feasible set* of the original problem. If the free variables are rewritten as above, and every inequality of the form "$\leq$" into an equation as in (18.4a) and (18.4b), then the set of all non-negative vectors $\underline{\mathbf{x}} \geq \underline{\mathbf{0}}$ satisfying the constraints is called the *feasible set* M:

$$M = \{\underline{\mathbf{x}} \in \mathbb{R}^n : \underline{\mathbf{x}} \geq \underline{\mathbf{0}},\ \mathbf{A}\underline{\mathbf{x}} = \underline{\mathbf{b}}\}. \tag{18.6a}$$

A point $\underline{\mathbf{x}}^* \in M$ with the property

$$f(\underline{\mathbf{x}}^*) \geq f(\underline{\mathbf{x}}) \quad \text{for every } \underline{\mathbf{x}} \in M \tag{18.6b}$$

is called the *maximum point* or the *solution point* of the linear programming problem. Obviously, the components of $\underline{\mathbf{x}}$ not belonging to slack variables form the solution of the original problem.

18.1.1.2 Examples and Graphical Solutions

1. Example of the Production of Two Products

Suppose primary materials R_1, R_2, and R_3 are needed to produce two products E_1 and E_2. **Scheme 18.1** shows how many units of primary materials are needed to produce each unit of the products E_1 and E_2, and there are given also the available amount of the primary materials.

Selling one unit of the products E_1 or E_2 results in 20 or 60 units of profit, respectively (PR). Determine a production program which yields maximum profit, if at least 10 units must be produced from product E_1.

Denoting by x_1 and x_2 the number of units produced from E_1 and E_2, the problem is:

Scheme 18.1	R_1 / E_i	R_2 / E_i	R_3 / E_i
E_1	12	8	0
E_2	6	12	10
Amount	630	620	350

$$\textbf{OF:} \quad f(\underline{\mathbf{x}}) = 20x_1 + 60x_2 = \max!$$

$$\textbf{CT:} \quad \begin{array}{rcl} 12x_1 + 6x_2 &\leq& 630, \\ 8x_1 + 12x_2 &\leq& 620, \\ 10x_2 &\leq& 350, \\ \\ x_1 \qquad\quad &\geq& 10. \end{array}$$

Introducing the slack variables x_3, x_4, x_5, x_6, one gets:

$$\textbf{OF}: \quad f(\underline{\mathbf{x}}) = 20x_1 + 60x_2 + 0\cdot x_3 + 0\cdot x_4 + 0\cdot x_5 + 0\cdot x_6 = \max!$$

$$\begin{array}{llllllll} \textbf{CT}: & 12x_1 + & 6x_2 + & x_3 & & & & = 630, \\ & 8x_1 + & 12x_2 & & + x_4 & & & = 620, \\ & & 10x_2 & & & + x_5 & & = 350, \\ & -x_1 & & & & & + x_6 & = -10. \end{array}$$

2. Properties of a Linear Programming Problem

On the basis of this example, some properties of the linear programming problem can be demonstrated by graphical representation. Here the slack variables are not considered; only the original two variables are used.

a) A line $a_1x_1+a_2x_2 = b$ divides the x_1, x_2 plane into two half-planes. The points (x_1, x_2) satisfying the inequality $a_1x_1+a_2x_2 \le b$ are in one of these half-planes. The graphical representation of this set of points in a Cartesian coordinate system can be made by a line, and the half-plane containing the solutions of the inequalities is denoted by an arrow. The set of feasible solutions M, i.e., the set of points satisfying all inequalities is the intersection of these half-planes **(Fig. 18.1)**.

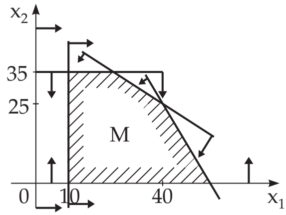


Figure 18.1

In this example the points of M form a polygonal domain. It may happen that M is unbounded or empty. If more then two boundary lines go through a vertex of the polygon, this vertex is called a degenerate vertex **(Fig. 18.2)**.

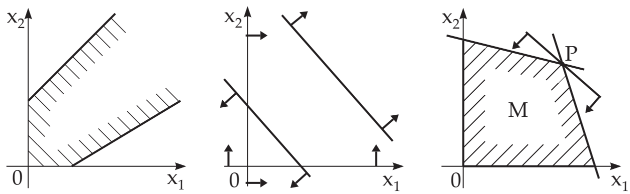


Figure 18.2

b) Every point in the x_1, x_2 plane satisfying the equality $f(x) = 20x_1 + 60x_2 = c_0$ is on one line, on the level line associated to the value c_0. With different choices of c_0, a family of parallel lines is defined, on each of which the value of the objective function is constant. Geometrically, those points are the solutions of the programming problem, which belong to the feasible set M and also to the level line $20x_1 + 60x_2 = c_0$ with maximal value of c_0. In this example, the solution point is $(x_1, x_2) = (25, 35)$ on the line $20x_1 + 60x_2 = 2600$. The level lines are represented in **Fig. 18.3**, where the arrows point in the direction of increasing values of the objective function.

Obviously, if the feasible set M is bounded, then there is at least one vertex such that the objective function takes its maximum. If the feasible set M is unbounded, it is possible that the objective function is unbounded, as well.

18.1.2 Basic Notions of Linear Programming, Normal Form

Now, the problem (18.5a,b) is considered with the feasible set M .

18.1.2.1 Extreme Points and Basis

1. Definition of the Extreme Point

A point $\underline{\mathbf{x}} \in M$ is called an *extreme point* or *vertex* of M, if for all $\underline{\mathbf{x}}_1, \underline{\mathbf{x}}_2 \in M$ with $\underline{\mathbf{x}}_1 \ne \underline{\mathbf{x}}_2$:

$$\underline{\mathbf{x}} \neq \lambda \underline{\mathbf{x}}_1 + (1-\lambda)\underline{\mathbf{x}}_2, \quad 0 < \lambda < 1, \tag{18.7}$$

i.e., $\underline{\mathbf{x}}$ is not on any line segment connecting two different points of M.

2. Theorem about Extreme Points

The point $\underline{\mathbf{x}} \in M$ is an *extreme point* of M if the columns of matrix $\mathbf{A}$ associated to the positive components of $\underline{\mathbf{x}}$ are linearly independent.
If the rank of $\mathbf{A}$ is m, then the maximal number of independent columns in $\mathbf{A}$ is m. So, an extreme point can have at most m positive components. The other components, at least $n-m$, are equal to zero. In the usual case, there are exactly m positive components. If the number of positive components is less then m, it is called a *degenerate extreme point*.

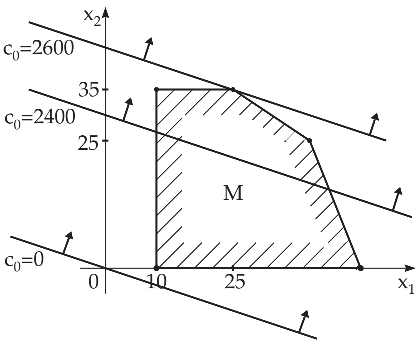


Figure 18.3

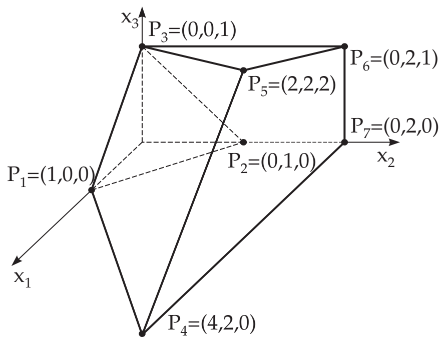


Figure 18.4

3. Basis

To every extreme point m linearly independent column vectors of the matrix $\mathbf{A}$ can be assigned, the columns belonging to the positive components. This system of linearly independent column vectors is called *the basis of the extreme point*. Usually, exactly one basis belongs to every extreme point. However several bases can be assigned to a degenerate extreme point. There are at most $\binom{n}{m}$ possibilities to choose m linearly independent vectors from n columns of $\mathbf{A}$. Consequently, the number of different bases, and therefore the number of different extreme points is $\binom{n}{m}$. If M is not empty, then M has at least one extreme point.

■ **OF:** $f(\underline{\mathbf{x}}) = 2x_1 + 3x_2 + 4x_3 = \max!$ **CT:**

$$\begin{aligned} x_1 + x_2 + x_3 &\geq 1,\\ x_2 &\leq 2,\\ -x_1 + 2x_3 &\leq 2,\\ 2x_1 - 3x_2 + 2x_3 &\leq 2. \end{aligned} \tag{18.8}$$

The feasible set M determined by the constraints is represented in **Fig. 18.4**. Introduction of slack variables x_4, x_5, x_6, x_7 leads to:

CT:
$$\begin{aligned} x_1 + x_2 + x_3 - x_4 &= 1,\\ x_2 + x_5 &= 2,\\ -x_1 + 2x_3 + x_6 &= 2,\\ 2x_1 - 3x_2 + 2x_3 + x_7 &= 2. \end{aligned}$$

The extreme point $P_2 = (0,1,0)$ of the polyhedron corresponds to the point $\underline{\mathbf{x}} = (x_1, x_2, x_3, x_4, x_5, x_6, x_7) = (0,1,0,0,1,2,5)$ of the extended system. The columns 2, 5, 6 and 7 of $\mathbf{A}$ form the corresponding basis. The degenerated extreme point P_1 corresponds to $(1,0,0,0,2,3,0)$. A basis of this extreme point contains the columns 1, 5, 6 and one of the columns 2, 4 or 7.

Remark: Here, the first inequality was a "$\geq$" inequality and x_4 was not added but subtracted. Frequently these types of additional variables both with a negative sign and a corresponding $b_i > 0$ are called *surplus variables*, rather than slack variables. As will be seen in 18.1.3.3, p. 916, the occurrence of surplus variables requires additional effort in the solution procedure.

4. Extreme Point with a Maximal Value of the Objective Function

Theorem: If M is not empty, and the objective function $f(\underline{\mathbf{x}}) = \underline{\mathbf{c}}^{\mathrm{T}}\underline{\mathbf{x}}$ is bounded from above on M, then there is at least one extreme point of M where it has its maximum.
A linear programming problem can be solved by determining at least one of the extreme points with maximum value of the objective function. Usually, the number of extreme points of M is very large in practical problems, so a method is needed by which the solution can be found in a reasonable time. Such a method is the *simplex method* which is also called the *simplex algorithm* or *simplex procedure.*

18.1.2.2 Normal Form of the Linear Programming Problem

1. Normal Form and Basic Solution

The linear programming problem (18.4a,b) can always be transformed to the following form with a suitable renumbering of the variables:

OF: $$f(\underline{\mathbf{x}}) = c_1x_1 + \cdots + c_{n-m}x_{n-m} + c_0 = \max! \tag{18.9a}$$

CT: $$\left.\begin{array}{l} a_{1,1}x_1 + \cdots + a_{1,n-m}x_{n-m} + x_{n-m+1} \qquad\qquad = b_1\,, \\ \quad\vdots \qquad\quad \vdots \qquad\qquad\qquad\qquad\qquad \ddots \qquad\quad \vdots \\ a_{m,1}x_1 + \cdots + a_{m,n-m}x_{n-m} \qquad\qquad\qquad + x_n = b_m\,, \\ \\ x_1, \ldots, x_{n-m}, x_{n-m+1}, \ldots, x_n \geq 0. \end{array}\right\} \tag{18.9b}$$

The last m columns of the coefficient matrix are obviously independent, and they form a basis. The *basic solution* $(x_1, x_2, \ldots, x_{n-m}, x_{n-m+1}, \ldots, x_n) = (0, \ldots, 0, b_1, \ldots, b_m)$ can be determined directly from the system of equations, but if $\underline{\mathbf{b}} \geq \underline{\mathbf{0}}$ does not hold, it is not a feasible solution.

If $\underline{\mathbf{b}} \geq \underline{\mathbf{0}}$, then (18.9a,b) is called a *normal form* or *canonical form of the linear programming problem*. In this case, the basic solution is a feasible solution, as well, i.e., $\underline{\mathbf{x}} \geq 0$, and it is an extreme point of M. The variables $x_1, \ldots, x_{n-m}$ are called *non-basic variables* and $x_{n-m+1}, \ldots, x_n$ are called *basic variables.* The objective function has the value c_0 at this extreme point, since the non-basic variables are equal to zero.

2. Determination of the Normal Form

If an extreme point of M is known, then a normal form of the linear programming problem (18.5a,b) can be obtained in the following way. A basis is chosen from the columns of $\mathbf{A}$ corresponding to the extreme point. Usually, these columns are determined by the positive components of the extreme point. Suppose the basic variables are collected into the vector $\underline{\mathbf{x}}_B$ and the non-basic variables are in $\underline{\mathbf{x}}_N$. The columns associated to the basis form the basis matrix $\mathbf{A}_B$, the other columns form the matrix $\mathbf{A}_N$. Then,

$$\mathbf{A}\underline{\mathbf{x}} = \mathbf{A}_N\underline{\mathbf{x}}_N + \mathbf{A}_B\underline{\mathbf{x}}_B = \underline{\mathbf{b}}. \tag{18.10}$$

The matrix $\mathbf{A}_B$ is non-singular and it has an inverse $\mathbf{A}_B^{-1}$, the so-called *basis inverse*. Multiplying (18.10) by $\mathbf{A}_B^{-1}$ and changing the objective function according to the non-basic variables results in the canonical form of the linear programming problem:

OF: $$f(\underline{\mathbf{x}}) = \underline{\mathbf{c}}_N^{\mathrm{T}}\underline{\mathbf{x}}_N + c_0, \tag{18.11a}$$

CT: $$\mathbf{A}_B^{-1}\mathbf{A}_N\underline{\mathbf{x}}_N + \underline{\mathbf{x}}_B = \mathbf{A}_B^{-1}\underline{\mathbf{b}} \quad \text{with} \quad \underline{\mathbf{x}}_N \geq \underline{\mathbf{0}},\ \underline{\mathbf{x}}_B \geq \underline{\mathbf{0}}. \tag{18.11b}$$

Remark: If the original system (18.1b) has only constraints of type "$\leq$" and simultaneously $b \geq \underline{\mathbf{0}}$, then the extended system (18.4b) contains no surplus variables (see 18.1.2.1, p. 911). In this case a normal form is immediately known. Selecting all slack variables as basic variables $\underline{\mathbf{x}}_B$ the result is $\mathbf{A}_B = \mathbf{I}$ and $\underline{\mathbf{x}}_B = \underline{\mathbf{b}}$ and $\underline{\mathbf{x}}_N = \underline{\mathbf{0}}$ is a feasible extreme point.

■ In the above example $\underline{\mathbf{x}} = (0,1,0,0,1,2,5)$ is an extreme point. Consequently:

$$\mathbf{A}_B = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -3 & 0 & 0 & 1 \end{pmatrix}, \quad \mathbf{A}_B^{-1} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 3 & 0 & 0 & 1 \end{pmatrix}, \quad \mathbf{A}_N = \begin{pmatrix} 1 & 1 & -1 \\ 0 & 0 & 0 \\ -1 & 2 & 0 \\ 2 & 2 & 0 \end{pmatrix}, \qquad (18.12a)$$

(columns of $\mathbf{A}_B$: $x_2\ x_5\ x_6\ x_7$; columns of $\mathbf{A}_N$: $x_1\ x_3\ x_4$)

$$\mathbf{A}_B^{-1}\mathbf{A}_N = \begin{pmatrix} 1 & 1 & -1 \\ -1 & -1 & 1 \\ -1 & 2 & 0 \\ 5 & 5 & -3 \end{pmatrix}, \qquad \mathbf{A}_B^{-1}\underline{\mathbf{b}} = \begin{pmatrix} 1 \\ 1 \\ 2 \\ 5 \end{pmatrix}. \qquad (18.12b)$$

(columns of $\mathbf{A}_B^{-1}\mathbf{A}_N$: $x_1\ x_3\ x_4$)

$$\left.\begin{array}{rl} x_1 + x_2 + x_3 - x_4 &= 1, \\ -x_1 - x_3 + x_4 + x_5 &= 1, \\ -x_1 + 2x_3 + x_6 &= 2, \\ 5x_1 + 5x_3 - 3x_4 + x_7 &= 5. \end{array}\right\} \qquad (18.13)$$

From $f(\underline{\mathbf{x}}) = 2x_1 + 3x_2 + 4x_3$ the transformed objective function

$$f(\underline{\mathbf{x}}) = -x_1 + x_3 + 3x_4 + 3 \qquad (18.14)$$

is obtained, if the triple of the first constraint is subtracted.

18.1.3 Simplex Method

18.1.3.1 Simplex Tableau

The *simplex method* is used to produce a sequence of extreme points of the feasible set with increasing values of the objective function. The transition to the new extreme point is performed starting from the normal form corresponding to the given extreme point, and arriving at the normal form corresponding to the new extreme point. In order to get a clear arrangement, and easier numerical performance, the normal form (18.9a,b) is represented in the simplex tableau **(Scheme 18.2a, 18.2b)**:

Scheme 18.2a

	x_1	$\cdots$	x_{n-m}	
x_{n-m+1}	$a_{1,1}$	$\cdots$	$a_{1,n-m}$	b_1
$\vdots$	$\vdots$		$\vdots$	$\vdots$
x_n	$a_{m,1}$	$\cdots$	$a_{m,n-m}$	b_m
	c_1	$\cdots$	c_{n-m}	$-c_0$

or briefly

Scheme 18.2b

	$\underline{\mathbf{x}}_N$	
$\underline{\mathbf{x}}_B$	$\mathbf{A}_N$	$\underline{\mathbf{b}}$
	$\underline{\mathbf{c}}$	$-c_0$

The k-th row of the tableau corresponds to the constraint

$$x_{n-m+k} + a_{k,1}x_1 + \cdots + a_{k,n-m}x_{n-m} = b_k. \qquad (18.15a)$$

The objective function is

$$c_1x_1 + \cdots + c_{n-m}x_{n-m} = f(\underline{\mathbf{x}}) - c_0. \qquad (18.15b)$$

From this simplex tableau, the extreme point $(\underline{\mathbf{x}}_N, \underline{\mathbf{x}}_B) = (\underline{\mathbf{0}}, \underline{\mathbf{b}})$ can be found. The value of the objective function at this point is $f(\underline{\mathbf{x}}) = c_0$. To put down $-c_0$ into the right below vertex of the tableau is advantageous for carrying out the simplex method. In every tableau always exactly one of the following three cases can be found:

a) $c_j \le 0,\ j = 1, \ldots, n-m$: The tableau is optimal. The point $(\underline{\mathbf{x}}_N, \underline{\mathbf{x}}_B) = (\underline{\mathbf{0}}, \underline{\mathbf{b}})$ is a maximal point. If all the c_j are positive, then this vertex is the only maximal point.

b) There exists at least one j such that $c_j > 0$ and $a_{ij} \le 0,\ i = 1, \ldots, m$: The linear programming problem has no solution, since the objective function is not bounded on the feasible set; for increasing

values of x_j it increases without a bound.

c) For every j with $c_j > 0$ there exists at least one i with $a_{ij} > 0$: It is possible to move from the extreme point $\underline{\mathbf{x}}$ to a neighboring extreme point $\tilde{\underline{\mathbf{x}}}$ with $f(\tilde{\underline{\mathbf{x}}}) \geq f(\underline{\mathbf{x}})$. In the case of a non-degenerate extreme point $\underline{\mathbf{x}}$, the ">" sign always holds.

18.1.3.2 Transition to the New Simplex Tableau

1. Non-Degenerate Case

If a tableau is not in final form (case **c)**), then a new tableau **(Scheme 18.3)** is determined. A basic variable x_p and a non-basic variable x_q are interchanged by the following calculations:

$$\textbf{a)}\ \tilde{a}_{pq} = \frac{1}{a_{pq}}. \tag{18.16a}$$

$$\textbf{b)}\ \tilde{a}_{pj} = a_{pj} \cdot \tilde{a}_{pq}, \quad j \neq q, \quad \tilde{b}_p = b_p \cdot \tilde{a}_{pq}. \tag{18.16b}$$

$$\textbf{c)}\ \tilde{a}_{iq} = -a_{iq} \cdot \tilde{a}_{pq}, \quad i \neq p, \quad \tilde{c}_q = -c_q \cdot \tilde{a}_{pq}. \tag{18.16c}$$

$$\textbf{d)}\ \tilde{a}_{ij} = a_{ij} + a_{pj} \cdot \tilde{a}_{iq}, \quad i \neq p,\ j \neq q,$$
$$\tilde{b}_i = b_i + b_p \cdot \tilde{a}_{iq}, \quad i \neq p, \quad \tilde{c}_j = c_j + a_{pj} \cdot \tilde{c}_q, \quad j \neq q, -\tilde{c}_0 = -c_0 + b_p \cdot \tilde{c}_q. \tag{18.16d}$$

The element a_{pq} is called the *pivot element* , the p-th row is the *pivot row*, and the q-th column is the *pivot column* . For the choice of a pivot element the following two requirements are to be considered:

a) $\tilde{c}_0 \geq c_0$ should hold;

b) the new tableau must also correspond to a feasible solution, i.e., $\tilde{\underline{\mathbf{b}}} \geq \underline{\mathbf{0}}$ must hold.

Then, $(\tilde{\underline{\mathbf{x}}}_N, \tilde{\underline{\mathbf{x}}}_B) = (\underline{\mathbf{0}}, \tilde{\underline{\mathbf{b}}})$ is a new extreme point at which the value of the objective function $f(\tilde{\underline{\mathbf{x}}}) = \tilde{c}_0$ is not smaller than it was previously. These conditions are satisfied if the pivot element is chosen in the following way:

a) To increase the value of the objective function, a column with $c_q > 0$ can be chosen for a pivot column;

b) to get a feasible solution, the pivot row must be chosen as

$$\frac{b_p}{a_{pq}} = \min_{\substack{1 \leq i \leq m \\ a_{iq} > 0}} \left\{ \frac{b_i}{a_{iq}} \right\}. \tag{18.17}$$

If the extreme points of the feasible set are not degenerate, then the simplex method terminates in a finite number of steps (case **a)** or case **b)**).

■ The normal form in 18.1.2, p. 911ff can be written in a simplex tableau **(Scheme 18.4a)**. This tableau is not optimal, since the objective function has a positive coefficient in the third column. The third column is assigned as the pivot column (the second column could also be taken under consideration). The quotients b_i/a_{iq} are calculated with every positive element of the pivot column (there is only one of them). The quotients are denoted behind the last column. The smallest quotient determines the pivot row.

Scheme 18.3

	$\tilde{\underline{\mathbf{x}}}_N$	
$\tilde{\underline{\mathbf{x}}}_B$	$\tilde{\mathbf{A}}_N$	$\tilde{\underline{\mathbf{b}}}$
	$\tilde{\underline{\mathbf{c}}}$	$-\tilde{c}_0$

Scheme 18.4a

	x_1	x_3	x_4		
x_2	1	1	$-\underline{1}$	1	
x_5	$-\underline{1}$	$-\underline{1}$	$\underline{\underline{1}}$	$\underline{1}$	1 : 1
x_6	−1	2	$\underline{0}$	2	
x_7	5	5	$-\underline{3}$	5	
	−1	1	$\underline{3}$	−3	

Scheme 18.4b

	x_1	x_3	x_5		
x_2	0	$\underline{0}$	1	2	
x_4	−1	$-\underline{1}$	1	1	
x_6	$-\underline{1}$	$\underline{\underline{2}}$	$\underline{0}$	$\underline{2}$	2 : 2
x_7	2	$\underline{2}$	3	8	8 : 2
	2	$\underline{4}$	−3	−6	

If it is not unique, then the extreme point corresponding to the new tableau is degenerate. After per-

forming the steps of (18.16a)–(18.16d) the tableau in **Scheme 18.4b** is obtained. This tableau determines the extreme point $(0, 2, 0, 1, 0, 2, 8)$, which corresponds to the point P_7 in **Fig. 18.4**. Since this new tableau is still not optimal, x_6 and x_3 are interchanged **(Scheme 18.4c)**. The extreme point of the third tableau corresponds to the point P_6 in **Fig. 18.4**. After an additional change an optimal tableau is obtained **(Scheme 18.4d)** with the maximal point $\underline{\mathbf{x}}^* = (2, 2, 2, 5, 0, 0, 0)$, which corresponds to the point P_5, and the objective function has a maximal value here: $f(\underline{\mathbf{x}}^*) = 18$.

Scheme 18.4c

	x_1	x_6	x_5		
x_2	0	0	1	2	
x_4	$-\frac{3}{2}$	$\frac{1}{2}$	1	2	
x_3	$-\frac{1}{2}$	$\frac{1}{2}$	0	1	
x_7	3	-1	3	6	$6:3$
	4	-2	-3	-10	

Scheme 18.4d

	x_7	x_6	x_5	
x_2	0	0	1	2
x_4	$\frac{1}{2}$	0	$\frac{5}{2}$	5
x_3	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{2}$	2
x_1	$\frac{1}{3}$	$-\frac{1}{3}$	1	2
	$-\frac{4}{3}$	$-\frac{2}{3}$	-7	-18

Scheme 18.5

	x_1	$\cdots$	x_n	
y_1	$a_{1,1}$	$\cdots$	$a_{1,n}$	b_1
$\vdots$	$\vdots$		$\vdots$	$\vdots$
y_m	$a_{m,1}$	$\cdots$	$a_{m,n}$	b_m
OF	c_1	$\cdots$	c_n	0
OF*	$\sum_{j=1}^{m} a_{j,1}$	$\cdots$	$\sum_{j=1}^{m} a_{j,n}$	$\sum_{j=1}^{m} b_j = -g(\underline{\mathbf{0}}, \underline{\mathbf{b}})$

2. Degenerate Case

If the next pivot element cannot be chosen uniquely in a simplex tableau, then the new tableau represents a degenerate extreme point. A degenerate extreme point can be interpreted geometrically as the coincident vertices of the convex polyhedron of the feasible solutions. There are several bases for such a vertex. In this case, it can therefore happen that some steps are performed without reaching a new extreme point. It is also possible that one gets a tableau that has occurred already before, so an infinite cycle may occur.

In the case of a degenerate extreme point, one possibility is to perturb the constants b_i by adding ε^i (with a suitable $\varepsilon^i > 0$) such that the resulting extreme points are no longer degenerate. The solution can be got from the solution of the perturbed problem, if $\varepsilon = 0$ is substituted.

If the pivot column is chosen "randomly" in the non-uniquely determined case, then the occurrence of an infinite cycle is unlikely in practical cases.

18.1.3.3 Determination of an Initial Simplex Tableau

1. Secondary Program, Artificial Variables

If there are equalities among the original constraints (18.1b) or inequalities with negative b_i, then it is not easy to find a feasible solution to start the simplex method. In this case, one starts with a secondary program to produce a feasible solution, which can be a starting point for a simplex procedure for the original problem. The system $\mathbf{A}\underline{\mathbf{x}} = \underline{\mathbf{b}}$ is modified by multiplying some of the equations with (-1) in order to satisfy the condition $\underline{\mathbf{b}} \geq \underline{\mathbf{0}}$. Now, an *artificial variable* $y_k \geq 0$ $(k = 1, 2, \ldots, m)$ is added to every left-hand side of $\mathbf{A}\underline{\mathbf{x}} = \underline{\mathbf{b}}$ with $\underline{\mathbf{b}} \geq \underline{\mathbf{0}}$, and the secondary program is considered:

$$\textbf{OF}^*\!: g(\underline{\mathbf{x}}, \underline{\mathbf{y}}) = -y_1 - \cdots - y_m = \max! \tag{18.18a}$$

$$\textbf{CT}^*\!: \left.\begin{array}{llllll} a_{1,1}x_1 & + \cdots + & a_{1,n}x_n & + y_1 & & = b_1, \\ \vdots & & \vdots & & \ddots & \vdots \\ a_{m,1}x_1 & + \cdots + & a_{m,n}x_n & & + y_m & = b_m, \\ x_1, \ldots, x_n \geq 0; & y_1, \ldots, y_m \geq 0. & & & & \end{array}\right\} \tag{18.18b}$$

For this problem, the variables $y_1, \dots, y_m$ are basic variables, and one can start the first simplex tableau **(Scheme 18.5)**. The last row of the tableau contains the sums of the coefficients of the non-basic variables , and these sums are the coefficients of the new secondary objective function $\mathbf{OF}^*$. Obviously, $g(\underline{\mathbf{x}}, \underline{\mathbf{y}}) \le 0$ always. If $g(\underline{\mathbf{x}}^*, \underline{\mathbf{y}}^*) = 0$ for a maximal point $(\underline{\mathbf{x}}^*, \underline{\mathbf{y}}^*)$ of the secondary problem, then obviously $\underline{\mathbf{y}}^* = 0$, and consequently $\underline{\mathbf{x}}^*$ is a solution of $\mathbf{A}\underline{\mathbf{x}} = \underline{\mathbf{b}}$. If $g(\underline{\mathbf{x}}^*, \underline{\mathbf{y}}^*) < 0$, then $\mathbf{A}\underline{\mathbf{x}} = \underline{\mathbf{b}}$ does not have any solution.

2. Solution of the Secondary Program

Our goal is to eliminate the artificial variables from the basis. A scheme is prepared not only for the secondary program separately. The original tableau is completed by columns of the artificial variables and the row of the secondary objective function. The secondary objective function now contains the sums of the corresponding coefficients from the rows corresponding to the equalities, as shown below. If an artificial variable becomes a non-basic variable, its column can be omitted, since it will be never chosen again as a basis variable. If a maximal point $(\underline{\mathbf{x}}^*, \underline{\mathbf{y}}^*)$ is determined, then two cases are distinguished:

1. $g(\underline{\mathbf{x}}^*, \underline{\mathbf{y}}^*) < 0$: The system $\mathbf{A}\underline{\mathbf{x}} = \underline{\mathbf{b}}$ has no solution, the linear programming problem does not have any feasible solution.

2. $g(\underline{\mathbf{x}}^*, \underline{\mathbf{y}}^*) = 0$: If there are no artificial variables among the basic variables, this tableau is an initial tableau for the original problem. Otherwise all artificial variables among the basic variables are removed by additional steps of the simplex method.

By introducing the artificial variables, the size of the problem can be increased considerably. It is not necessary to introduce artificial variables for every equation. If the system of constraints before introducing the slack and surplus variables (see Remark in 18.1.2, **3.**, p. 913) has the form $\mathbf{A}_1\underline{\mathbf{x}} \ge \underline{\mathbf{b}}_1$, $\mathbf{A}_2\underline{\mathbf{x}} = \underline{\mathbf{b}}_2$, $\mathbf{A}_3\underline{\mathbf{x}} \le \underline{\mathbf{b}}_3$ with $\underline{\mathbf{b}}_1, \underline{\mathbf{b}}_2, \underline{\mathbf{b}}_3 \ge 0$, then artificial variables must be introduced only for the first two systems. For the third system the slack variables can be chosen as basic variables.

■ In the example of 18.1.2, p. 912, only the first equation requires an artificial variable:

$$\textbf{OF}^*:\quad g(\underline{\mathbf{x}}, \underline{\mathbf{y}}) = \qquad -y_1 = \max!$$

$$\textbf{CT}^*:\quad \begin{aligned} x_1 + x_2 + x_3 - x_4 + y_1 &= 1,\\ x_2 + x_5 &= 2,\\ -x_1 + 2x_3 + x_6 &= 2,\\ 2x_1 - 3x_2 + 2x_3 + x_7 &= 2. \end{aligned}$$

The tableau **(Scheme 18.6b)** is optimal with $g(\underline{\mathbf{x}}^*, \underline{\mathbf{y}}^*) = 0$. After omitting the second column the first tableau of the original problem is obtained.

Scheme 18.6a

	x_1	x_2	x_3	x_4		
y_1	$\underline{1}$	$\underline{1}$	$\underline{1}$	$-\underline{1}$	$\underline{1}$	1 : 1
x_5	0	$\underline{1}$	0	0	2	2 : 1
x_6	−1	$\underline{0}$	2	0	2	
x_7	2	$-\underline{3}$	2	0	2	
OF	2	$\underline{3}$	4	0	0	
OF*	1	$\underline{1}$	1	−1	1	

Scheme 18.6b

	x_1	y_1	x_3	x_4	1
x_2	1	1	1	−1	1
x_5	−1	−1	−1	1	1
x_6	−1	0	2	0	2
x_7	5	3	5	−3	5
OF	−1	−3	1	3	−3
OF*	0	−1	0	0	0

18.1.3.4 Revised Simplex Method

1. Revised Simplex Tableau

Suppose the linear programming problem is given in normal form:

$$\textbf{OF:}\quad f(\underline{\mathbf{x}}) = c_1x_1 + \cdots + c_{n-m}x_{n-m} + c_0 = \max! \tag{18.19a}$$

CT:
$$\left.\begin{array}{llll} \alpha_{1,1}x_1 + \cdots + \alpha_{1,n-m}x_{n-m} + x_{n-m+1} & & & = \beta_1, \\ \vdots \qquad\qquad \vdots & & \ddots & \vdots \\ \alpha_{m,1}x_1 + \cdots + \alpha_{m,n-m}x_{n-m} & & + x_n & = \beta_m, \\ x_1 \ge 0, \ldots, x_n \ge 0. & & & \end{array}\right\} \quad (18.19b)$$

Obviously, the coefficient vectors $\underline{\alpha}_{n-m+i}$ $(i = 1, \ldots, n)$ are the i-th unit vectors.
In order to change into another normal form and therefore to reach another extreme point, it is sufficient to multiply the system of equations (18.19b) by the corresponding basis inverse. (Recall the fact that if $\mathbf{A}_B$ denotes a new basis, then the coordinates of a vector $\underline{\mathbf{x}}$ can be expressed in this new basis as $\mathbf{A}_B^{-1}\underline{\mathbf{x}}$. If the inverse of the new basis is known, then any column as well as the objective function from the very first tableau can be got by simple multiplication.) The simplex method can be modified so that only the basis inverse is determined in every step instead of a new tableau. From every tableau only those elements are calculated which are required to find the new pivot element. If the number of variables is considerably larger than the number of constraints ($n > 3m$), then the revised simplex method requires considerably less computing cost and therefore has better accuracy.
The general form of a revised simplex tableau is shown in **Scheme 18.7**.

Scheme 18.7

	$x_1 \cdots x_{n-m}$	$x_{n-m+1} \cdots x_n$		x_q
x_1^B		$a_{1,n-m+1} \cdots a_{1,n}$	b_1	r_1
$\vdots$		$\vdots \qquad \vdots$	$\vdots$	$\vdots$
x_m^B		$a_{m,n-m+1} \cdots a_{m,n}$	b_m	r_m
	$c_1 \cdots c_{n-m}$	$c_{n-m+1} \cdots c_n$	$-c_0$	c_q

The quantities of the scheme have the following meaning:

$x_1^B, \ldots, x_m^B$: Actual basic variables (in the first step the same as $x_{n-m+1} \cdots x_n$).
$c_1, \ldots, c_n$: Coefficients of the objective function (the coefficients associated to the basic variables are zeros).
$b_1, \ldots, b_m$: Right-hand side of the actual normal form.
c_0 : Value of the objective function at the extreme point $(x_1^B, \ldots, x_m^B) = (b_1, \ldots, b_m)$.

$$\mathbf{A}^* = \begin{pmatrix} a_{1,n-m+1} & \cdots & a_{1,n} \\ \vdots & & \vdots \\ a_{m,n-m+1} & \cdots & a_{m,n} \end{pmatrix}:$$
Actual basis inverse, where the columns of $\mathbf{A}^*$ are the columns of $x_{n-m+1}, \ldots, x_n$ corresponding to the actual normal form;

$\underline{\mathbf{r}} = (r_1, \ldots, r_m)^{\mathrm{T}}$: Actual pivot column.

2. Revised Simplex Step

a) The tableau is not optimal when at least one of the coefficients c_j $(j = 1, 2, \ldots, n)$ is positiv. A pivot column q is chosen for a $c_q > 0$.
b) One calculates the pivot column $\underline{\mathbf{r}}$ by multiplying the q-th column of the original coefficient matrix (18.19b) by $\mathbf{A}^*$ and introduces the new vector as the last vector of the tableau.
The pivot row k is determined in the same way as in the simplex algorithm (18.17).
c) The new tableau is calculated by the pivoting step (18.16a–d), where a_{iq} is formally replaced by r_i and the indices are restricted for $n - m + 1 \le j \le n$. The column $\tilde{\underline{\mathbf{r}}}$ is omitted. x_q becomes a basic variable. For $j = 1, \ldots, n - m$, the results are $\tilde{c}_j = c_j + \underline{\alpha}_j^{\mathrm{T}}\tilde{\underline{\mathbf{c}}}$, where $\tilde{\underline{\mathbf{c}}} = (\tilde{c}_{n-m+1}, \ldots, \tilde{c}_n)^{\mathrm{T}}$, and $\underline{\alpha}_j$ is the j-th column of the coefficient matrix of (18.19b).

■ Consider the normal form of the example in 18.1.2, p. 912. One wants to bring x_4 into the basis. The corresponding pivot column $\underline{\mathbf{r}} = \underline{\alpha}_4$ is placed into the last column of the tableau **(Scheme 18.8a)** (initially $\mathbf{A}^*$ is the unit matrix).

For $j = 1, 3, 4$ one gets $\tilde{c}_j = c_j - 3\alpha_{2j}$: $(c_1, c_3, c_4) = (2, 4, 0)$.
The determined extreme point $\underline{\mathbf{x}} = (0, 2, 0, 1, 0, 2, 8)$ corresponds to the point P_7 in **Fig. 18.4**, p. 912.
The next pivot column can be chosen for $j = 3 = q$.

Scheme 18.8a

	x_1	x_3	x_4	x_2	x_5	x_6	x_7		x_4	
x_2				1	0	0	0	1	-1	
x_5				$\underline{0}$	$\underline{1}$	$\underline{0}$	$\underline{0}$	$\underline{1}$	$\underline{\underline{1}}$	$1:1$
x_6				0	0	1	0	2	0	
x_7				0	0	0	1	5	-3	
	-1	1	$\underline{3}$	0	0	0	0	-3	$\underline{3}$	

Scheme 18.8b

	x_1	x_3	x_4	x_2	x_5	x_6	x_7		x_3	
x_2				1	1	0	0	2	0	
x_4				0	1	0	0	1	-1	
x_6				$\underline{0}$	$\underline{0}$	$\underline{1}$	$\underline{0}$	$\underline{2}$	$\underline{\underline{2}}$	$2:2$
x_7				0	3	0	1	8	2	$8:2$
	2	$\underline{4}$	-3	0	-3	0	0	-6	$\underline{4}$	

The vector $\underline{\mathbf{r}}$ is determined by

$$\underline{\mathbf{r}} = (r_1, \ldots, r_m) = \mathbf{A}^* \underline{\alpha}_3 = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 3 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ -1 \\ 2 \\ 5 \end{pmatrix} = \begin{pmatrix} 0 \\ -1 \\ 2 \\ 2 \end{pmatrix}$$

and it is placed into the very last column of the second tableau **(Scheme 18.8b)**. One proceedes as above analogously to the method shown in 18.1.3.2, p. 915. If one wants to return to the original method, then the matrix of the original columns of the non-basic variables must be multiplied by $\mathbf{A}^*$ and only these columns will be kept.

18.1.3.5 Duality in Linear Programming

1. Correspondence

To any linear programming problem (primal problem) an other unique linear programming problem can be assigned (dual problem):

Primal problem

OF: $f(\underline{\mathbf{x}}) = \underline{\mathbf{c}}_1^{\mathrm{T}} \underline{\mathbf{x}}_1 + \underline{\mathbf{c}}_2^{\mathrm{T}} \underline{\mathbf{x}}_2 = \max!$ (18.20a)

CT:
$$\begin{aligned} \mathbf{A}_{1,1}\underline{\mathbf{x}}_1 + \mathbf{A}_{1,2}\underline{\mathbf{x}}_2 &\le \underline{\mathbf{b}}_1, \\ \mathbf{A}_{2,1}\underline{\mathbf{x}}_1 + \mathbf{A}_{2,2}\underline{\mathbf{x}}_2 &= \underline{\mathbf{b}}_2, \\ \underline{\mathbf{x}}_1 \ge 0, \quad \underline{\mathbf{x}}_2 &\text{ free.} \end{aligned} \tag{18.20b}$$

Dual problem

OF*: $g(\underline{\mathbf{u}}) = \underline{\mathbf{b}}_1^{\mathrm{T}} \underline{\mathbf{u}}_1 + \underline{\mathbf{b}}_2^{\mathrm{T}} \underline{\mathbf{u}}_2 = \min!$ (18.21a)

CT*:
$$\begin{aligned} \mathbf{A}_{1,1}^{\mathrm{T}}\underline{\mathbf{u}}_1 + \mathbf{A}_{2,1}^{\mathrm{T}}\underline{\mathbf{u}}_2 &\ge \underline{\mathbf{c}}_1, \\ \mathbf{A}_{1,2}^{\mathrm{T}}\underline{\mathbf{u}}_1 + \mathbf{A}_{2,2}^{\mathrm{T}}\underline{\mathbf{u}}_2 &= \underline{\mathbf{c}}_2, \\ \underline{\mathbf{u}}_1 \ge 0, \quad \underline{\mathbf{u}}_2 &\text{ free.} \end{aligned} \tag{18.21b}$$

The coefficients of the objective function of one of the problems form the right-hand side vector of the constraints of the other problem. Every free variable corresponds to an equation, and every variable with restricted sign corresponds to an inequality of the other problem.

2. Duality Theorems

a) If both problems have feasible solutions, i.e., $M \ne \emptyset$, $M^* \ne \emptyset$ (where M and M^* denote the feasible sets of the primal and dual problems respectively), then

$$f(\underline{\mathbf{x}}) \le g(\underline{\mathbf{u}}) \quad \text{for all } \underline{\mathbf{x}} \in M, \ \underline{\mathbf{u}} \in M^*, \tag{18.22a}$$

and both problems have optimal solutions.

b) The points $\underline{\mathbf{x}} \in M$ and $\underline{\mathbf{u}} \in M^*$ are optimal solutions for the corresponding problem, if and only if

$$f(\underline{\mathbf{x}}) = g(\underline{\mathbf{u}}). \tag{18.22b}$$

c) If $f(\underline{\mathbf{x}})$ has no upper bound on M or $g(\underline{\mathbf{u}})$ has no lower bound on M^*, then $M^* = \emptyset$ or $M = \emptyset$, i.e., the dual problem has no feasible solution.

d) The points $\underline{\mathbf{x}} \in M$ and $\underline{\mathbf{u}} \in M^*$ are optimal points of the corresponding problems if and only if:

$$\underline{\mathbf{u}}_1^{\mathrm{T}}(\mathbf{A}_{1,1}\underline{\mathbf{x}}_1 + \mathbf{A}_{1,2}\underline{\mathbf{x}}_2 - \underline{\mathbf{b}}_1) = 0 \quad \text{and} \quad \underline{\mathbf{x}}_1^{\mathrm{T}}(\mathbf{A}_{1,1}^{\mathrm{T}}\underline{\mathbf{u}}_1 + \mathbf{A}_{2,1}^{\mathrm{T}}\underline{\mathbf{u}}_2 - \underline{\mathbf{c}}_1) = 0. \tag{18.22c}$$

Using these last equations, a solution $\underline{\mathbf{x}}$ of the primal problem can be found from a non-degenerate optimal solution $\underline{\mathbf{u}}$ of the dual problem by solving the following linear system of equations:

$$\mathbf{A}_{2,1}\underline{\mathbf{x}}_1 + \mathbf{A}_{2,2}\underline{\mathbf{x}}_2 - \underline{\mathbf{b}}_2 = \underline{\mathbf{0}}, \tag{18.23a}$$

$$(\mathbf{A}_{1,1}\underline{\mathbf{x}}_1 + \mathbf{A}_{1,2}\underline{\mathbf{x}}_2 - \underline{\mathbf{b}}_1)_i = \underline{\mathbf{0}} \quad \text{for } u_i > 0, \tag{18.23b}$$

$$x_i = 0 \quad \text{for } (\mathbf{A}_{1,1}^{\mathrm{T}}\underline{\mathbf{u}}_1 + \mathbf{A}_{2,1}^{\mathrm{T}}\underline{\mathbf{u}}_2 - \underline{\mathbf{c}}_1)_i \neq 0. \tag{18.23c}$$

The dual problem also can be solved by the simplex method.

3. Application of the Dual Problem

Working with the dual problem may have some advantages in the following cases:

a) If it is simple to find a normal form for the dual problem, one switches from the primal problem to the dual.

b) If the primal problem has a large number of constraints compared to the number of variables, then the revised simplex method can be used for the dual problem.

■ Consider the original problem of the example of 18.1.2, p. 912.

Primal problem

OF: $f(\underline{\mathbf{x}}) = 2x_1 + 3x_2 + 4x_3 = \max!$

CT:
$$\begin{array}{rl} -x_1 - x_2 - x_3 & \le -1, \\ x_2 & \le 2, \\ -x_1 + 2x_3 & \le 2, \\ 2x_1 - 3x_2 + 2x_3 & \le 2, \\ x_1, x_2, x_3 & \ge 0. \end{array}$$

Dual problem

OF*: $g(\underline{\mathbf{u}}) = -u_1 + 2u_2 + 2u_3 + 2u_4 = \min!$

CT*:
$$\begin{array}{rl} -u_1 - u_3 + 2u_4 & \ge 2, \\ -u_1 + u_2 - 3u_4 & \ge 3, \\ -u_1 + 2u_3 + 2u_4 & \ge 4, \\ u_1, u_2, u_3, u_4 & \ge 0. \end{array}$$

If the dual problem is solved by the simplex method after introducing the slack variables, then the optimal solution $\underline{\mathbf{u}}^* = (u_1, u_2, u_3, u_4) = (0, 7, 2/3, 4/3)$ with $g(\underline{\mathbf{u}}) = 18$ is got. A solution $\underline{\mathbf{x}}^*$ of the primal problem can be got by solving the system $(\mathbf{A}\underline{\mathbf{x}} - \underline{\mathbf{b}})_i = 0$ for $u_i > 0$, i.e., $x_2 = 2$, $-x_1 + 2x_3 = 2$, $2x_1 - 3x_2 + 2x_3 = 2$, therefore: $\underline{\mathbf{x}}^* = (2, 2, 2)$ with $f(\underline{\mathbf{x}}) = 18$.

18.1.4 Special Linear Programming Problems

18.1.4.1 Transportation Problem

1. Modeling

A certain product, produced by m producers $E_1, E_2, \ldots, E_m$ in quantities $a_1, a_2, \ldots, a_m$, is to be transported to n consumers $V_1, V_2, \ldots, V_n$ with demands $b_1, b_2, \ldots, b_n$. Transportation cost of a unit product of producer E_i to consumer V_j is c_{ij}. The amount of the product transported from E_i to V_j is x_{ij} units. An optimal transportation plan is to be determined with minimum total transportation cost. The system is supposed to be balanced, i.e., supply equals demand:

$$\sum_{i=1}^{m} a_i = \sum_{j=1}^{n} b_j. \tag{18.24}$$

The matrix of costs $\mathbf{C}$ and the distribution matrix $\mathbf{X}$ are constructed:

$$\mathbf{C} = \begin{pmatrix} c_{1,1} & \cdots & c_{1,n} \\ \vdots & & \vdots \\ c_{m,1} & \cdots & c_{m,n} \end{pmatrix} \begin{matrix} E: \\ E_1 \\ \vdots \\ E_m \end{matrix}, \quad V: \; V_1 \cdots V_n \tag{18.25a}$$

$$\mathbf{X} = \begin{pmatrix} x_{1,1} & \cdots & x_{1,n} \\ \vdots & & \vdots \\ x_{m,1} & \cdots & x_{m,n} \end{pmatrix} \begin{matrix} \Sigma: \\ a_1 \\ \vdots \\ a_m \end{matrix}. \quad \Sigma: \; b_1 \cdots b_n \tag{18.25b}$$

If condition (18.24) is not fulfilled, then two cases are distinguished:

a) If $\sum a_i > \sum b_j$, then a fictitious consumer V_{n+1} is introduced with demand $b_{n+1} = \sum a_i - \sum b_j$ and with transportation costs $c_{i,n+1} = 0$.

b) If $\sum a_i < \sum b_j$, then introduce a fictitious producer E_{m+1} is introduced with capacity $a_{m+1} =$

$\sum b_j - \sum a_i$ and with transportation costs $c_{m+1,j} = 0$.

In order to determine an optimal program, the following programming problem should be solved:

$$\textbf{OF}: \quad f(\mathbf{X}) = \sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij} x_{ij} = \min! \tag{18.26a}$$

$$\textbf{CT}: \quad \sum_{j=1}^{n} x_{ij} = a_i \;\; (i = 1, \ldots, m), \quad \sum_{i=1}^{m} x_{ij} = b_j \;\; (j = 1, \ldots, n), \;\; x_{ij} \geq 0). \tag{18.26b}$$

The minimum of the problem occurs at a vertex of the feasible set. There are $m + n - 1$ linearly independent constraints among the $m + n$ original constraints, so, in the non-degenerate case, the solution contains $m + n - 1$ positive components x_{ij}. To determine an optimal solution the following algorithm is used, which is called the transportation algorithm.

2. Determination of a Basic Feasible Solution

With the *Northwest corner rule* an initial basic feasible solution can be determined:

a) Choose $\quad x_{11} = \min(a_1, b_1)$. (18.27a)

b) If $a_1 < b_1$, the first row of $\mathbf{X}$ is omitted. (18.27b)

If $a_1 > b_1$, the first column of $\mathbf{X}$ is omitted. (18.27c)

If $a_1 = b_1$, either the first row or the first remaining column of $\mathbf{X}$ is omitted. (18.27d)

If there are only one row but several columns, then one column is cancelled. The same applies for the rows.

c) a_1 is replaced by $a_1 - x_{11}$ and b_1 by $b_1 - x_{11}$ and the procedure in repeated in the left upper vertex of the reduced distribution matrix $\mathbf{X}$.

The variables obtained in step **a)** are the basic variables, all the others are non-basic variables with zero values.

■

$$\mathbf{C} = \begin{array}{c} \\ \left(\begin{array}{cccc} 5 & 3 & 2 & 7 \\ 8 & 2 & 1 & 1 \\ 9 & 2 & 6 & 3 \end{array}\right) \\ V: \; V_1 \; V_2 \; V_3 \; V_4 \end{array} \begin{array}{l} E: \\ E_1 \\ E_2\,, \\ E_3 \\ \\ \end{array} \qquad \mathbf{X} = \begin{array}{c} \\ \left(\begin{array}{cccc} x_{1,1} & x_{1,2} & x_{1,3} & x_{1,4} \\ x_{2,1} & x_{2,2} & x_{2,3} & x_{2,4} \\ x_{3,1} & x_{3,2} & x_{3,3} & x_{3,4} \end{array}\right) \\ \sum: \; b_1 = 4 \; b_2 = 6 \; b_3 = 5 \; b_4 = 7 \end{array} \begin{array}{l} \sum: \\ a_1 = 9 \\ a_2 = 10\,. \\ a_3 = 3 \\ \\ \end{array}$$

The determination of an initial extreme point with the Northwest corner rule gives

first step   second step   further steps

$$\mathbf{X} = \begin{array}{c} \left(\begin{array}{cccc} 4 & & & \\ | & & & \\ | & & & \end{array}\right) \\ \not{4}\; 6\; 5\; 7 \\ 0 \end{array} \begin{array}{l} \not{9}\;\; 5 \\ 10 \\ 3 \\ \\ \\ \end{array}, \qquad \mathbf{X} = \begin{array}{c} \left(\begin{array}{cccc} 4 & 5 & \multicolumn{2}{c}{\text{—}} \\ | & & & \\ | & & & \end{array}\right) \\ 0\; \not{6}\; 5\; 7 \\ \not{1} \end{array} \begin{array}{l} \not{5}\;\; 0 \\ 10 \\ 3 \\ \\ \\ \end{array}, \qquad \mathbf{X} = \begin{array}{c} \left(\begin{array}{cccc} 4 & 5 & \multicolumn{2}{c}{\text{—}} \\ | & 1 & 5 & 4 \\ | & | & | & 3 \end{array}\right) \\ 0 \quad \not{5}\; \not{7} \\ \not{1}\; 0\; 3 \\ 0 \end{array} \begin{array}{l} \qquad 0 \\ 1\not{0}\; \not{9}\; \not{4}\; 0\,. \\ 3 \\ \\ \\ \end{array}$$

There are alternative methods to find an initial basic solution which also takes the transportation costs into consideration (see, e.g., the Vogel approximation method in [18.13]) and they usually result in a better initial solution.

3. Solution of the Transportation Problem with the Simplex Method

If the usual simplex tableau is prepared for this problem, then it results in a huge tableau $((m+n) \times (m \cdot n))$ with a large number of zeros: In each column, only two elements are equal to 1. So, a reduced tableau is constructed, and the following steps correspond to the simplex steps working only with the non-zero elements of the theoretical simplex tableau. The matrix of the cost data contains the coefficients of

the objective function. The basic variables are exchanged for non-basic variables iteratively, while the corresponding elements of the cost matrix are modified in each step. The procedure is explained by an example.

a) Determination of the modified cost matrix $\tilde{\mathbf{C}}$ from $\mathbf{C}$ by

$$\tilde{c}_{ij} = c_{ij} + p_i + q_j \quad (i = 1, \ldots, m, \;\; j = 1, \ldots, n), \tag{18.28a}$$

with the conditions

$$\tilde{c}_{ij} = 0 \;\text{ for }\; (i, j) \;\text{ if } x_{ij} \text{ is an actual basic variable.} \tag{18.28b}$$

The elements of $\mathbf{C}$ belonging to basic variables are marked and $p_1 = 0$ is substituted. The other quantities p_i and q_j, also called potentials or simplex multiplicators, are determined so that the sum of p_i, q_j and the marked costs c_{ij} should be 0:

■

$$\mathbf{C} = \begin{pmatrix} (5) & (3) & 2 & 7 \\ 8 & (2) & (1) & (1) \\ 9 & 2 & 6 & (3) \end{pmatrix} \begin{matrix} p_1 = 0 \\ p_2 = 1 \\ p_3 = -1 \end{matrix} \quad \Longrightarrow \quad \tilde{\mathbf{C}} = \begin{pmatrix} 0 & 0 & 0 & 5 \\ 4 & 0 & 0 & 0 \\ 3 & \boxed{-2} & 3 & 0 \end{pmatrix}. \tag{18.28c}$$
$$q_1 = -5 \;\; q_2 = -3 \;\; q_3 = -2 \;\; q_4 = -2$$

b) The value

$$\tilde{c}_{pq} = \min_{i,j}\{\tilde{c}_{ij}\} \tag{18.28d}$$

must be determined. If $\tilde{c}_{pq} \geq 0$, then the given distribution $\mathbf{X}$ is optimal; otherwise x_{pq} is chosen as a new basic variable. In our example: $\tilde{c}_{pq} = \tilde{c}_{32} = -2$.

c) In $\tilde{\mathbf{C}}$, $\tilde{c}_{pq}$ and the costs associated to the basic variables are marked. If $\tilde{\mathbf{C}}$ contains rows or columns with at most one marked element, then these rows or columns will be omitted. This procedure is repeated with the remaining matrix, until no further cancellation is possible.

$$\tilde{\mathbf{C}} = \begin{pmatrix} (0) & (0) & 0 & 5 \\ 4 & (0) & (0) & (0) \\ 3 & (-2) & 3 & (0) \end{pmatrix}. \tag{18.28e}$$

d) The elements x_{ij} associated to the remaining marked elements $\tilde{c}_{ij}$ form a cycle. The new basic variable $\tilde{x}_{pq}$ is to be set to a positive value δ. The other variables $\tilde{x}_{ij}$ associated to the marked elements $\tilde{c}_{ij}$ are determined by the constraints. In practice, δ is subtracted and added from or to every second element of the cycle. To keep the variables non-negative, the amount δ must be chosen as

$$\delta = x_{rs} = \min\{x_{ij}\colon \;\; \tilde{x}_{ij} = x_{ij} - \delta\}, \tag{18.28f}$$

where x_{rs} will be the non-basic variable. In the example $\delta = \min\{1, 3\} = 1$.

$$\tilde{\mathbf{X}} = \begin{pmatrix} 4 & 5 & & \\ & \longleftarrow & & \\ & 1-\delta & 5 & 4+\delta \\ & \downarrow & & \uparrow \\ & \delta & \longrightarrow & 3-\delta \end{pmatrix} \begin{matrix} \Sigma \\ 9 \\ \\ 10 \\ \\ 3 \end{matrix} \quad \Longrightarrow \quad \tilde{\mathbf{X}} = \begin{pmatrix} 4 & 5 & & \\ & & 5 & 5 \\ & 1 & & 2 \end{pmatrix}, \quad f(\underline{\mathbf{x}}) = 53. \tag{18.28g}$$
$$\Sigma \;\; 4 \;\; 6 \;\; 5 \;\; 7$$

Then, this procedure is repeated with $\mathbf{X} = \tilde{\mathbf{X}}$.

$$\mathbf{C} = \begin{pmatrix} (5) & (3) & 2 & 7 \\ 8 & 2 & (1) & (1) \\ 9 & (2) & 6 & (3) \end{pmatrix} \begin{matrix} p_1 = 0 \\ p_2 = 3 \\ p_3 = 1 \end{matrix} \quad \Longrightarrow \quad \tilde{\mathbf{C}} = \begin{pmatrix} (0) & (0) & (\underline{-2}) & 3 \\ 6 & 2 & (0) & (0) \\ 5 & (0) & 3 & (0) \end{pmatrix}, \tag{18.28h}$$
$$q_1 = -5 \;\; q_2 = -3 \;\; q_3 = -4 \;\; q_4 = -4$$

$$\tilde{\mathbf{X}} = \begin{pmatrix} 4 & 5-\delta \leftarrow & \delta & \\ & & \uparrow & \\ & \downarrow & 5-\delta \leftarrow & 5+\delta \\ & & & \uparrow \\ & 1+\delta & \longrightarrow & 2-\delta \end{pmatrix} \quad \overset{\delta=2}{\Longrightarrow} \quad \tilde{\mathbf{X}} = \begin{pmatrix} 4 & 3 & 2 & \\ & & 3 & 7 \\ & 3 & & \end{pmatrix}, \quad f(\mathbf{X}) = 49. \tag{18.28i}$$

The next matrix $\tilde{\mathbf{C}}$ does not contain any negative element. So, $\tilde{\mathbf{X}}$ is an optimal solution.

18.1.4.2 Assignment Problem

The representation is made by an example.

■ n shipping contracts should be given to n shipping companies so that each company receives exactly one contract. The assignment has to be determined which minimizes the total costs, if the i-th company charges c_{ij} for the j-th contract.

An assignment problem is a special transportation problem with $m = n$ and $a_i = b_j = 1$ for all i, j:

$$\textbf{OF}: \quad f(\underline{\mathbf{x}}) = \sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}x_{ij} = \min! \tag{18.29a}$$

$$\textbf{CT}: \quad \sum_{j=1}^{n} x_{ij} = 1 \;\; (i = 1, \ldots, n), \qquad \sum_{i=1}^{n} x_{ij} = 1 \;\; (j = 1, \ldots, n), \qquad x_{ij} \in \{0, 1\}. \tag{18.29b}$$

Every feasible distribution matrix contains exactly one 1 in every row and every column, all other elements are equal to zero. In a general transportation problem of this dimension, however, a non-degenerate basic solution would have $2n - 1$ positive variables. Thus, basic feasible solutions to the assignment problem are highly degenerate, with $n - 1$ basic variables equal to zero. Starting with a feasible distribution matrix $\mathbf{X}$, the assignment problem can be solved by the general transportation algorithm. It is time consuming to do so. However, because of the highly degenerate nature of the basic feasible solutions, the assignment problem can be solved with the highly efficient *Hungarian method* (see [18.9]).

18.1.4.3 Distribution Problem

The problem is represented by an example.

■ m products $E_1, E_2, \ldots, E_m$ should be produced in quantities $a_1, a_2, \ldots, a_m$. Every product can be produced on any of n machines $M_1, M_2, \ldots, M_n$. The production of a unit of product E_i on machine M_j needs processing time b_{ij} and cost c_{ij}. The time capacity of machine M_j is b_j. Denote the quantity produced by machine M_j from product E_i by x_{ij}. The total production costs should be minimized.

This distribution problem has the following general model:

$$\textbf{OF}: \quad f(\underline{\mathbf{x}}) = \sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij}x_{ij} = \min! \tag{18.30a}$$

$$\textbf{CT}: \quad \sum_{j=1}^{m} x_{ij} = a_i \;\; (i = 1, \ldots, m), \quad \sum_{i=1}^{n} b_{ij}x_{ij} \le b_j \;\; (j = 1, \ldots, n), \quad x_{ij} \ge 0 \;\text{ for all }\; i, j. \tag{18.30b}$$

The distribution problem is a generalization of the transportation problem and it can be solved by the simplex method. If all $b_{ij} = 1$, then the more effective transportation algorithm can be used (see 18.1.4.1, p. 921) after introducing a fictitious product E_{m+1} (see 18.1.4.1, p. 920).

18.1.4.4 Travelling Salesman

Suppose there are n places $O_1, O_2, \ldots, O_n$. The travelling time from O_i to O_j is c_{ij}. Here, $c_{ij} \ne c_{ji}$ is possible.

One wants to determine the shortest route such that the traveller passes through every place exactly once, and returns to the starting point.

Similarly to the assignment problem, exactly one element is chosen in every row and column of the time

matrix $\mathbf{C}$ so that the sum of the chosen elements is minimal. The difficulty of the numerical solution of this problem is the restriction that the marked elements c_{ij} should be arranged in order of the following form:

$$c_{i_1,i_2}, c_{i_2,i_3}, \ldots, c_{i_n,i_{n+1}} \quad \text{with} \;\; i_k \neq i_l \;\; \text{for} \;\; k \neq l \;\; \text{and} \;\; i_{n+1} = i_1. \tag{18.31}$$

The travelling salesman problem can be solved by the branch and bound methods.

18.1.4.5 Scheduling Problem

n different products are processed on m different machines in a product-dependent order. At any time only one product can be processed on a machine. The processing time of each product on each machine is assumed to be known. Waiting times, when a given product is not in process, and machine idle times are also possible.
An optimal scheduling of the processing jobs is determined where the objective function is selected as the time when all jobs are finished, or the total waiting time of jobs, or total machine idle time. Sometimes the sum of the finishing times for all jobs is chosen as the objective function when no waiting time or idle time is allowed.

18.2 Non-linear Optimization

18.2.1 Formulation of the Problem, Theoretical Basis

18.2.1.1 Formulation of the Problem

1. Non-linear Optimization Problem
A non-linear optimization problem has the general form

$$f(\underline{\mathbf{x}}) = \min! \quad \text{subject to} \quad \underline{\mathbf{x}} \in \mathbb{R}^n \quad \text{with} \tag{18.32a}$$

$$g_i(\underline{\mathbf{x}}) \leq 0, \qquad i \in I = \{1, \ldots, m\}, \quad h_j(\underline{\mathbf{x}}) = 0, \quad j \in J = \{1, \ldots, r\} \tag{18.32b}$$

where at least one of the functions f, g_i, h_j is non-linear. The set of feasible solutions is denoted by

$$M = \{\underline{\mathbf{x}} \in \mathbb{R}^n : \;\; g_i(\underline{\mathbf{x}}) \leq 0, \quad i \in I, \quad h_j(\underline{\mathbf{x}}) = 0, \quad j \in J\}. \tag{18.33}$$

The problem is to determine the minimum points.

2. Minimum Points
A point $\underline{\mathbf{x}}^* \in M$ is called the *global minimum point* if $f(\underline{\mathbf{x}}^*) \leq f(\underline{\mathbf{x}})$ holds for every $\underline{\mathbf{x}} \in M$. If this relation holds for only the points $\underline{\mathbf{x}}$ of a neighborhood U of $\underline{\mathbf{x}}^*$, then $\underline{\mathbf{x}}^*$ is called a *local minimum point.*
Since the equality constraints $h_j(\underline{\mathbf{x}}) = 0$ can be expressed by two inequalities,

$$-h_j(\underline{\mathbf{x}}) \leq 0, \quad h_j(\underline{\mathbf{x}}) \leq 0, \tag{18.34}$$

it can be supposed that the set J is empty, $J = \emptyset$.

18.2.1.2 Optimality Conditions

1. Special Directions
a) The Cone of the Feasible Directions at $\underline{\mathbf{x}} \in M$ is defined by

$$Z(\underline{\mathbf{x}}) = \{\underline{\mathbf{d}} \in \mathbb{R}^n : \;\; \exists \bar{\alpha} > 0 : \;\; \underline{\mathbf{x}} + \alpha\underline{\mathbf{d}} \in M, \quad 0 \leq \alpha \leq \bar{\alpha}\}, \quad \underline{\mathbf{x}} \in M, \tag{18.35}$$

where the directions are denoted by $\underline{\mathbf{d}}$. If $\underline{\mathbf{d}} \in Z(\underline{\mathbf{x}})$, then every point of the ray $\underline{\mathbf{x}} + \alpha\underline{\mathbf{d}}$ belongs to M for sufficient small values of α.

b) A Descent Direction at a point $\underline{\mathbf{x}}$ is a vector $\underline{\mathbf{d}} \in \mathbb{R}^n$ for which there exists an $\bar{\alpha} > 0$ such that

$$f(\underline{\mathbf{x}} + \alpha\underline{\mathbf{d}}) < f(\underline{\mathbf{x}}) \quad \forall \alpha \in (0, \bar{\alpha}). \tag{18.36}$$

There exists no feasible descent direction at a minimum point.
If f is differentiable, then $\underline{\mathbf{d}}$ is a descent direction when $\nabla f(\underline{\mathbf{x}})^{\mathrm{T}}\underline{\mathbf{d}} < 0$. Here, ∇ denotes the nabla operator, so $\nabla f(\underline{\mathbf{x}})$ represents the gradient of the scalar-valued function f at $\underline{\mathbf{x}}$.

2. Necessary Optimality Conditions
If f is differentiable and $\underline{\mathbf{x}}^*$ is a local minimum point, then

$$\nabla f(\underline{\mathbf{x}}^*)^{\mathrm{T}}\underline{\mathbf{d}} \geq 0 \quad \text{for every} \quad \underline{\mathbf{d}} \in \overline{Z}(\underline{\mathbf{x}}^*). \tag{18.37a}$$

In particular, if $\underline{\mathbf{x}}^*$ is an interior point of M, then

$$\nabla f(\underline{\mathbf{x}}^*) = \underline{\mathbf{0}}. \tag{18.37b}$$

3. Lagrange Function and Saddle Point

Optimality conditions (18.37a,b) should be transformed into a more practical form including the constraints. The so-called Lagrange function or *Lagrangian* is constructed:

$$L(\underline{\mathbf{x}}, \underline{\mathbf{u}}) = f(\underline{\mathbf{x}}) + \sum_{i=1}^{m} u_i g_i(\underline{\mathbf{x}}) = f(\underline{\mathbf{x}}) + \underline{\mathbf{u}}^{\mathrm{T}} g(\underline{\mathbf{x}}), \quad \underline{\mathbf{x}} \in \mathbb{R},\ \underline{\mathbf{u}} \in \mathbb{R}^m_+, \tag{18.38}$$

according to the *Lagrange multiplier method* (see 6.2.5.6, p. 456) for problems with equality constraints. A point $(\underline{\mathbf{x}}^*, \underline{\mathbf{u}}^*) \in \mathbb{R}^n \times \mathbb{R}^m_+$ is called a *saddle point* of L, if

$$L(\underline{\mathbf{x}}^*, \underline{\mathbf{u}}) \le L(\underline{\mathbf{x}}^*, \underline{\mathbf{u}}^*) \le L(\underline{\mathbf{x}}, \underline{\mathbf{u}}^*) \quad \text{for every } \underline{\mathbf{x}} \in \mathbb{R}^n,\ \underline{\mathbf{u}} \in \mathbb{R}^m_+. \tag{18.39}$$

4. Global Kuhn-Tucker Conditions

A point $\underline{\mathbf{x}}^* \in \mathbb{R}^n$ satisfies the *global Kuhn-Tucker conditions* if there is an $\underline{\mathbf{u}}^* \in \mathbb{R}^m_+$, i.e., $\underline{\mathbf{u}}^* \ge 0$ such that $(\underline{\mathbf{x}}^*, \underline{\mathbf{u}}^*)$ is a saddle point of L.
For the proof of the Kuhn-Tucker conditions see 12.5.6, p. 683.

5. Sufficient Optimality Condition

If $(\underline{\mathbf{x}}^*, \underline{\mathbf{u}}^*) \in \mathbb{R}^n \times \mathbb{R}^m_+$ is a saddle point of L, then $\underline{\mathbf{x}}^*$ is a global minimum point of (18.32a,b).
If the functions f and g_i are differentiable, then local optimality conditions can be deduced.

6. Local Kuhn-Tucker Conditions

A point $\underline{\mathbf{x}}^* \in M$ satisfies the local Kuhn–Tucker conditions if there are numbers $u_i \ge 0$, $i \in I_0(\underline{\mathbf{x}}^*)$ such that

$$-\nabla f(\underline{\mathbf{x}}^*) = \sum_{i \in I_0(\underline{\mathbf{x}}^*)} u_i \nabla g_i(\underline{\mathbf{x}}^*), \quad \text{where} \tag{18.40a}$$

$$I_0(\underline{\mathbf{x}}) = \{i \in \{1, \dots, m\} \ :\ g_i(\underline{\mathbf{x}}) = 0\} \tag{18.40b}$$

is the index set of the active constraints at $\underline{\mathbf{x}}$. The point $\underline{\mathbf{x}}^*$ is also called a *Kuhn-Tucker stationary point.*

This means geometrically that a point $\underline{\mathbf{x}}^* \in M$ satisfies the local Kuhn-Tucker conditions, if the negative gradient $-\nabla f(\underline{\mathbf{x}}^*)$ lies in the cone spanned by the gradients $\nabla g_i(\underline{\mathbf{x}}^*)$ $i \in I_0(\underline{\mathbf{x}}^*)$ of the constraints active at $\underline{\mathbf{x}}^*$ **(Fig. 18.5)**.
The following equivalent formulation for (18.40a,b) is also often used: $\underline{\mathbf{x}}^* \in \mathbb{R}^n$ satisfies the local Kuhn-Tucker conditions, if there is a $\underline{\mathbf{u}}^* \in \mathbb{R}^m_+$ such that

$$g(\underline{\mathbf{x}}^*) \le 0, \tag{18.41a}$$

$$u_i g_i(\underline{\mathbf{x}}^*) = 0, \quad i = 1, \dots, m, \tag{18.41b}$$

$$\nabla f(\underline{\mathbf{x}}^*) + \sum_{i=1}^{m} u_i \nabla g_i(\underline{\mathbf{x}}^*) = 0. \tag{18.41c}$$

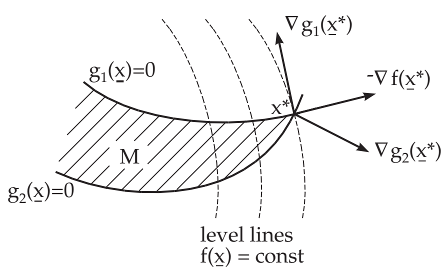


Figure 18.5

7. Necessary Optimality Conditions and Kuhn-Tucker Conditions

If $\underline{\mathbf{x}}^* \in M$ is a local minimum point of (18.32a,b) and the feasible set satisfies the *regularity condition* at $\underline{\mathbf{x}}^* :\ \exists\, \underline{\mathbf{d}} \in \mathbb{R}^n$ such that $\nabla g_i(\underline{\mathbf{x}}^*)^{\mathrm{T}} \underline{\mathbf{d}} < 0$ for every $i \in I_0(\underline{\mathbf{x}}^*)$, then $\underline{\mathbf{x}}^*$ satisfies the local Kuhn-Tucker conditions.

18.2.1.3 Duality in Optimization

1. Dual Problem

With the associated Lagrangian (18.38) the maximum problem is formed, the so-called dual of (18.32a,b):

$$L(\underline{\mathbf{x}}, \underline{\mathbf{u}}) = \max! \quad \text{subject to} \quad (\underline{\mathbf{x}}, \underline{\mathbf{u}}) \in M^* \quad \text{with} \tag{18.42a}$$

$$M^* = \{(\underline{\mathbf{x}}, \underline{\mathbf{u}}) \in \mathbb{R}^n \times \mathbb{R}^m_+ \;:\; L(\underline{\mathbf{x}}, \underline{\mathbf{u}}) = \min_{\underline{\mathbf{z}} \in \mathbb{R}^n} L(\underline{\mathbf{z}}, \underline{\mathbf{u}})\}. \tag{18.42b}$$

2. Duality Theorems

If $\underline{\mathbf{x}}_1 \in M$ and $(\underline{\mathbf{x}}_2, \underline{\mathbf{u}}_2) \in M^*$, then

a) $L(\underline{\mathbf{x}}_2, \underline{\mathbf{u}}_2) \le f(\underline{\mathbf{x}}_1)$.

b) If $L(\underline{\mathbf{x}}_2, \underline{\mathbf{u}}_2) = f(\underline{\mathbf{x}}_1)$, then $\underline{\mathbf{x}}_1$ is a minimum point of (18.32a,b) and $(\underline{\mathbf{x}}_2, \underline{\mathbf{u}}_2)$ is a maximum point of (18.42a,b).

18.2.2 Special Non-linear Optimization Problems

18.2.2.1 Convex Optimization

1. Convex Problem

The optimization problem

$$f(\underline{\mathbf{x}}) = \min! \quad \text{subject to} \quad g_i(\underline{\mathbf{x}}) \le 0 \quad (i = 1, \ldots, m) \tag{18.43}$$

is called a *convex problem* if the functions f and g_i are convex. In particular, f and g_i can be linear functions. The following statements are valid for convex problems:

a) Every local minimum of f over M is also a global minimum.

b) If M is not empty and bounded, then there exists at least one solution of (18.43).

c) If f is strictly convex, then there is at most one solution of (18.43).

1. Optimality Conditions

a) If f has continuous partial derivatives, then $\underline{\mathbf{x}}^* \in M$ is a solution of (18.43), if

$$(\underline{\mathbf{x}} - \underline{\mathbf{x}}^*)^{\mathrm{T}} \nabla f(\underline{\mathbf{x}}^*) \ge 0 \quad \text{for every} \quad \underline{\mathbf{x}} \in M. \tag{18.44}$$

b) The *Slater condition* is a regularity condition for the feasible set M. It is satisfied if there exists an $\underline{\mathbf{x}} \in M$ such that $g_i(\underline{\mathbf{x}}) < 0$ for every non-affine linear functions g_i.

c) If the Slater condition is satisfied, then $\underline{\mathbf{x}}^*$ is a minimum point of (18.43) if and only if there exists a $\underline{\mathbf{u}}^* \ge 0$ such that $(\underline{\mathbf{x}}^*, \underline{\mathbf{u}}^*)$ is a saddle point of the Lagrangian. Moreover, if functions f and g_i are differentiable, then $\underline{\mathbf{x}}^*$ is a solution of (18.43) if and only if $\underline{\mathbf{x}}^*$ satisfies the local Kuhn-Tucker conditions.

d) The dual problem (18.42a,b) can be formulated easily for a convex optimization problem with differentiable functions f and g_i:

$$L(\underline{\mathbf{x}}, \underline{\mathbf{u}}) = \max!, \quad \text{subject to} \quad (\underline{\mathbf{x}}, \underline{\mathbf{u}}) \in M^* \quad \text{with} \tag{18.45a}$$

$$M^* = \{(\underline{\mathbf{x}}, \underline{\mathbf{u}}) \in \mathbb{R}^n \times \mathbb{R}^m_+ \;:\; \nabla_{\underline{\mathbf{x}}} L(\underline{\mathbf{x}}, \underline{\mathbf{u}}) = \underline{\mathbf{0}}\}. \tag{18.45b}$$

The gradient of L is calculated here only with respect to $\underline{\mathbf{x}}$.

e) For convex optimization problems, the *strong duality theorem* also holds:
If M satisfies the Slater condition and if $\underline{\mathbf{x}}^* \in M$ is a solution of (18.43), then there exists a $\underline{\mathbf{u}}^* \in \mathbb{R}^m_+$, such that $(\underline{\mathbf{x}}^*, \underline{\mathbf{u}}^*)$ is a solution of the dual problem (18.45a,b), and

$$f(\underline{\mathbf{x}}^*) = \min_{\underline{\mathbf{x}} \in M} f(\underline{\mathbf{x}}) = \max_{(\underline{\mathbf{x}}, \underline{\mathbf{u}}) \in M^*} L(\underline{\mathbf{x}}, \underline{\mathbf{u}}) = L(\underline{\mathbf{x}}^*, \underline{\mathbf{u}}^*). \tag{18.46}$$

18.2.2.2 Quadratic Optimization

1. Formulation of the Problem

Quadratic optimization problems have the form

$$f(\underline{\mathbf{x}}) = \underline{\mathbf{x}}^{\mathrm{T}} \mathbf{C} \underline{\mathbf{x}} + \underline{\mathbf{p}}^{\mathrm{T}} \underline{\mathbf{x}} = \min!, \quad \text{subject to} \quad \underline{\mathbf{x}} \in M \subset \mathbb{R}^n \quad \text{with} \tag{18.47a}$$

$$M = M_I: \quad M = \{\underline{\mathbf{x}} \in \mathbb{R}^n \;:\; \mathbf{A}\underline{\mathbf{x}} \le \underline{\mathbf{b}},\ \underline{\mathbf{x}} \ge \underline{\mathbf{0}}\}. \tag{18.47b}$$

Here, $\mathbf{C}$ is a symmetric (n,n) matrix, $\underline{\mathbf{p}} \in \mathbb{R}^n$, $\mathbf{A}$ is an (m,n) matrix, and $\underline{\mathbf{b}} \in \mathbb{R}^m$.
The feasible set M can be written alternatively in the following way:

$$M = M_{II}: \quad M = \{\underline{\mathbf{x}} \ : \ \mathbf{A}\underline{\mathbf{x}} = \underline{\mathbf{b}},\ \underline{\mathbf{x}} \geq \underline{\mathbf{0}}\}, \tag{18.48a}$$

$$M = M_{III}: \quad M = \{\underline{\mathbf{x}} \ : \ \mathbf{A}\underline{\mathbf{x}} \leq \underline{\mathbf{b}}\}. \tag{18.48b}$$

2. Lagrangian and Kuhn-Tucker Conditions

The Lagrangian to the problem (18.47a,b) is

$$L(\underline{\mathbf{x}}, \underline{\mathbf{u}}) = \underline{\mathbf{x}}^{\mathrm{T}}\mathbf{C}\underline{\mathbf{x}} + \underline{\mathbf{p}}^{\mathrm{T}}\underline{\mathbf{x}} + \underline{\mathbf{u}}^{\mathrm{T}}(\mathbf{A}\underline{\mathbf{x}} - \underline{\mathbf{b}}). \tag{18.49}$$

By introducing the notation

$$\underline{\mathbf{v}} = \frac{\partial L}{\partial \underline{\mathbf{x}}} = \underline{\mathbf{p}} + 2\mathbf{C}\underline{\mathbf{x}} + \mathbf{A}^{\mathrm{T}}\underline{\mathbf{u}} \quad \text{and} \quad \underline{\mathbf{y}} = -\frac{\partial L}{\partial \underline{\mathbf{u}}} = -\mathbf{A}\underline{\mathbf{x}} + \underline{\mathbf{b}} \tag{18.50}$$

the Kuhn-Tucker conditions are as follows:

	Case I:	Case II:	Case III:	
a)	$\mathbf{A}\underline{\mathbf{x}} + \underline{\mathbf{y}} = \underline{\mathbf{b}}$,	$\mathbf{A}\underline{\mathbf{x}} = \underline{\mathbf{b}}$,	$\mathbf{A}\underline{\mathbf{x}} + \underline{\mathbf{y}} = \underline{\mathbf{b}}$,	(18.51a)
b)	$2\mathbf{C}\underline{\mathbf{x}} - \underline{\mathbf{v}} + \mathbf{A}^{\mathrm{T}}\underline{\mathbf{u}} = -\underline{\mathbf{p}}$,	$2\mathbf{C}\underline{\mathbf{x}} - \underline{\mathbf{v}} + \mathbf{A}^{\mathrm{T}}\underline{\mathbf{u}} = -\underline{\mathbf{p}}$,	$2\mathbf{C}\underline{\mathbf{x}} + \mathbf{A}^{\mathrm{T}}\underline{\mathbf{u}} = -\underline{\mathbf{p}}$,	(18.51b)
c)	$\underline{\mathbf{x}} \geq \underline{\mathbf{0}},\ \underline{\mathbf{v}} \geq \underline{\mathbf{0}},\ \underline{\mathbf{y}} \geq \underline{\mathbf{0}},\ \underline{\mathbf{u}} \geq \underline{\mathbf{0}}$,	$\underline{\mathbf{x}} \geq \underline{\mathbf{0}},\ \underline{\mathbf{v}} \geq \underline{\mathbf{0}}$,	$\underline{\mathbf{u}} \geq \underline{\mathbf{0}},\ \underline{\mathbf{y}} \geq \underline{\mathbf{0}}$,	(18.51c)
d)	$\underline{\mathbf{x}}^{\mathrm{T}}\underline{\mathbf{v}} + \underline{\mathbf{y}}^{\mathrm{T}}\underline{\mathbf{u}} = 0$.	$\underline{\mathbf{x}}^{\mathrm{T}}\underline{\mathbf{v}} = 0$.	$\underline{\mathbf{y}}^{\mathrm{T}}\underline{\mathbf{u}} = 0$.	(18.51d)

3. Convexity

The function $f(\underline{\mathbf{x}})$ is convex (strictly convex) if and only if the matrix $\mathbf{C}$ is positive semidefinite (positive definite). Every result on convex optimization problems can be used for quadratic problems with a positive semidefinite matrix $\mathbf{C}$; in particular, the Slater condition always holds, so it is necessary and sufficient for the optimality of a point $\underline{\mathbf{x}}^*$ that there exists a point $(\underline{\mathbf{x}}^*, \underline{\mathbf{y}}, \underline{\mathbf{u}}, \underline{\mathbf{v}})$, which satisfies the corresponding system of local Kuhn-Tucker conditions.

4. Dual Problem

If $\mathbf{C}$ is positive definite, then the dual problem (18.45a,b) of (18.47a,b) can be expressed explicitly:

$$L(\underline{\mathbf{x}}, \underline{\mathbf{u}}) = \max!, \quad \text{subject to} \quad (\underline{\mathbf{x}}, \underline{\mathbf{u}}) \in M^*, \quad \text{where} \tag{18.52a}$$

$$M^* = \{(\underline{\mathbf{x}}, \underline{\mathbf{u}}) \in \mathbb{R}^n \times \mathbb{R}^m_+ \ : \ \underline{\mathbf{x}} = -\frac{1}{2}\mathbf{C}^{-1}(\mathbf{A}^{\mathrm{T}}\underline{\mathbf{u}} + \underline{\mathbf{p}})\}. \tag{18.52b}$$

If the expression $\underline{\mathbf{x}} = -\frac{1}{2}\mathbf{C}^{-1}(\mathbf{A}^{\mathrm{T}}\underline{\mathbf{u}} + \underline{\mathbf{p}})$ is substituted into the dual objective function $L(\underline{\mathbf{x}}, \underline{\mathbf{u}})$, then the equivalent problem is

$$\varphi(\underline{\mathbf{u}}) = -\frac{1}{4}\underline{\mathbf{u}}^{\mathrm{T}}\mathbf{A}\mathbf{C}^{-1}\mathbf{A}^{\mathrm{T}}\underline{\mathbf{u}} - \left(\frac{1}{2}\mathbf{A}\mathbf{C}^{-1}\underline{\mathbf{p}} + \underline{\mathbf{b}}\right)^{\mathrm{T}}\underline{\mathbf{u}} - \frac{1}{4}\underline{\mathbf{p}}^{\mathrm{T}}\underline{\mathbf{C}}^{-1}\underline{\mathbf{p}} = \max!, \quad \underline{\mathbf{u}} \geq \underline{\mathbf{0}}. \tag{18.53}$$

Hence: If $\underline{\mathbf{x}}^* \in M$ is a solution of (18.47a,b), then (18.53) has a solution $\underline{\mathbf{u}}^* \geq \underline{\mathbf{0}}$, and

$$f(\underline{\mathbf{x}}^*) = \varphi(\underline{\mathbf{u}}^*). \tag{18.54}$$

Problem (18.53) can be replaced by an equivalent formulation:

$$\psi(\underline{\mathbf{u}}) = \underline{\mathbf{u}}^{\mathrm{T}}\mathbf{E}\underline{\mathbf{u}} + \underline{\mathbf{h}}^{\mathrm{T}}\underline{\mathbf{u}} = \min!\,, \quad \text{subject to} \quad \underline{\mathbf{u}} \geq \underline{\mathbf{0}} \quad \text{where} \tag{18.55a}$$

$$\mathbf{E} = \frac{1}{4}\mathbf{A}\mathbf{C}^{-1}\mathbf{A}^{\mathrm{T}} \quad \text{and} \quad \underline{\mathbf{h}} = \frac{1}{2}\mathbf{A}\mathbf{C}^{-1}\underline{\mathbf{p}} + \underline{\mathbf{b}}\,. \tag{18.55b}$$

18.2.3 Solution Methods for Quadratic Optimization Problems

18.2.3.1 Wolfe's Method

1. Formulation of the Problem and Solution Principle

The method of Wolfe is to solve quadratic problems of the special form:

$$f(\underline{\mathbf{x}}) = \underline{\mathbf{x}}^{\mathrm{T}}\mathbf{C}\underline{\mathbf{x}} + \underline{\mathbf{p}}^{\mathrm{T}}\underline{\mathbf{x}} = \min!\,, \quad \text{subject to} \quad \mathbf{A}\underline{\mathbf{x}} = \underline{\mathbf{b}}, \quad \underline{\mathbf{x}} \geq \underline{\mathbf{0}}. \tag{18.56}$$

$\mathbf{C}$ is supposed to be positive definite. The basic idea is the determination of a solution $(\underline{\mathbf{x}}^*, \underline{\mathbf{u}}^*, \underline{\mathbf{v}}^*)$ of the corresponding system of Kuhn-Tucker conditions, associated to problem (18.56):

$$\mathbf{A}\underline{\mathbf{x}} = \underline{\mathbf{b}}, \tag{18.57a}$$

$$2\mathbf{C}\underline{\mathbf{x}} - \underline{\mathbf{v}} + \mathbf{A}^{\mathrm{T}}\underline{\mathbf{u}} = -\underline{\mathbf{p}}, \tag{18.57b}$$

$$\underline{\mathbf{x}} \geq \underline{\mathbf{0}}, \quad \underline{\mathbf{v}} \geq \underline{\mathbf{0}}; \tag{18.57c}$$

$$\underline{\mathbf{x}}^{\mathrm{T}}\underline{\mathbf{v}} = 0. \tag{18.58}$$

Relations (18.57a,b,c) represent a linear equation system with $m + n$ equations and $2n + m$ variables. Because of relation (18.58), either $x_i = 0$ or $v_i = 0$ $(i = 1, 2, \dots, n)$ must hold. Therefore, every solution of (18.57a,b,c), (18.58) contains at most $m+n$ non-zero components. Hence, it must be a basic solution of (18.57a,b,c).

2. Solution Process

First, a feasible basic solution (vertex) $\bar{\underline{\mathbf{x}}}$ of the system $\mathbf{A}\underline{\mathbf{x}} = \underline{\mathbf{b}}$ is determined. The indices belonging to the basis variables of $\bar{\underline{\mathbf{x}}}$ form the set I_B. In order to find a solution of system (18.57a,b,c), which also satisfies (18.58), the problem is formulated as

$$-\mu = \min!\,, \quad (\mu \in \mathbb{R}); \tag{18.59}$$

$$\mathbf{A}\underline{\mathbf{x}} = \underline{\mathbf{b}}, \tag{18.60a}$$

$$2\mathbf{C}\underline{\mathbf{x}} - \underline{\mathbf{v}} + \mathbf{A}^{\mathrm{T}}\underline{\mathbf{u}} - \mu\underline{\mathbf{q}} = -\underline{\mathbf{p}} \quad \text{with} \quad \underline{\mathbf{q}} = 2\mathbf{C}\bar{\underline{\mathbf{x}}} + \underline{\mathbf{p}}, \tag{18.60b}$$

$$\underline{\mathbf{x}} \geq \underline{\mathbf{0}}, \ \underline{\mathbf{v}} \geq \underline{\mathbf{0}}, \ \mu \geq 0; \tag{18.60c}$$

$$\underline{\mathbf{x}}^{\mathrm{T}}\underline{\mathbf{v}} = 0. \tag{18.61}$$

If $(\underline{\mathbf{x}}, \underline{\mathbf{v}}, \underline{\mathbf{u}}, \mu)$ is a solution of this problem also satisfying (18.57a,b,c) and (18.58), then $\mu = 0$.
The vector $(\underline{\mathbf{x}}, \underline{\mathbf{v}}, \underline{\mathbf{u}}, \mu) = (\bar{\underline{\mathbf{x}}}, \underline{\mathbf{0}}, \underline{\mathbf{0}}, 1)$ is a known feasible solution of the system (18.60a,b,c), and it satisfies the relation (18.61), too. A basis associated to this basic solution is formed from the columns of the coefficient matrix

$$\begin{pmatrix} \mathbf{A} & \mathbf{0} & \mathbf{0} & \underline{\mathbf{0}} \\ 2\mathbf{C} & -\mathbf{I} & \mathbf{A}^{\mathrm{T}} & -\underline{\mathbf{q}} \end{pmatrix}, \tag{18.62}$$

$\mathbf{I}$ denotes the unit matrix, $\mathbf{0}$ the zero matrix and $\underline{\mathbf{0}}$ is the zero vector of the corresponding dimension,

in the following way:
a) m columns belonging to x_i with $i \in I_B$,
b) $n - m$ columns belonging to v_i with $i \notin I_B$,
c) all m columns belonging to u_i,
d) the last column, but then a suitable column determined in **b)** or **c)** will be dropped.
If $\underline{\mathbf{q}} = \underline{\mathbf{0}}$, then the interchange according to **d)** is not possible. Then $\bar{\underline{\mathbf{x}}}$ is already a solution.
Now, a first simplex tableau can be constructed. The minimization of the objective function is performed by the simplex method with an additional rule that guarantees that the relation $\underline{\mathbf{x}}^{\mathrm{T}}\underline{\mathbf{v}} = 0$ is satisfied:
The variables x_i and v_i $(i = 1, 2, \dots, n)$ must not be simultaneously basic variables.
In the case of a positive definite $\mathbf{C}$, considering this additional rule the simplex method provides a solution of problem (18.59), (18.60a,b,c), (18.61) satisfying $\mu = 0$. For a positive semi-definite matrix $\mathbf{C}$, because of the restricted pivot choice, it may happen that although $\mu > 0$, no more exchange-step can be made without violating the additional rules. In this case μ cannot be reduced any further.

■ $f(\underline{\mathbf{x}}) = x_1^2 + 4x_2^2 - 10x_1 - 32x_2 = \min!$ with $x_1 + 2x_2 + x_3 = 7, \quad 2x_1 + x_2 + x_4 = 8.$

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 1 & 0 \\ 2 & 1 & 0 & 1 \end{pmatrix}, \quad \underline{\mathbf{b}} = \begin{pmatrix} 7 \\ 8 \end{pmatrix}, \quad \mathbf{C} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad \underline{\mathbf{p}} = \begin{pmatrix} -10 \\ -32 \\ 0 \\ 0 \end{pmatrix}.$$

In this case $\mathbf{C}$ is positive semi-definite. A feasible basic solution of $\mathbf{A}\underline{\mathbf{x}} = \underline{\mathbf{b}}$ is $\bar{\underline{\mathbf{x}}} = (0, 0, 7, 8)^{\mathrm{T}}$, $\underline{\mathbf{q}} = 2\mathbf{C}\bar{\underline{\mathbf{x}}} + \underline{\mathbf{p}} = (-10, -32, 0, 0)^{\mathrm{T}}$. The choices for the basis vectors are: **a)** columns 3 and 4 of $\begin{pmatrix} \mathbf{A} \\ 2\mathbf{C} \end{pmatrix}$, **b)** columns 1 and 2 of $\begin{pmatrix} \mathbf{0} \\ -\mathbf{I} \end{pmatrix}$, **c)** the columns of $\begin{pmatrix} \mathbf{0} \\ \mathbf{A}^{\mathrm{T}} \end{pmatrix}$ and **d)** column $\begin{pmatrix} \mathbf{0} \\ -\underline{\mathbf{q}} \end{pmatrix}$ instead of the first column of $\begin{pmatrix} \mathbf{0} \\ -\mathbf{I} \end{pmatrix}$. The basis matrix is formed from these columns, and the basis inverse is calculated (see 18.1, p. 909). Multiplying matrix (18.62) and the vectors $\begin{pmatrix} \underline{\mathbf{b}} \\ -\underline{\mathbf{p}} \end{pmatrix}$ by the basis inverse, the first simplex tableau **(Scheme 18.9)** is obtained. Only x_1 can be interchanged with v_2 in this tableau according to the complementary constraints. After a few steps, we get the solution $\underline{\mathbf{x}}^* = (2, 5/2, 0, 3/2)^{\mathrm{T}}$ is obtained. The last two equations of $2\mathbf{C}\underline{\mathbf{x}} - \underline{\mathbf{v}} + \mathbf{A}^{\mathrm{T}}\underline{\mathbf{u}} - \mu\underline{\mathbf{q}} = -\underline{\mathbf{p}}$ are: $v_3 = u_1,\ v_4 = u_2$. Therefore, by eliminating u_1 and u_2 the dimension of the problem can be reduced.

Scheme 18.9

	x_1	x_2	v_1	v_3	v_4	
x_3	1	2	0	0	0	7
x_4	2	1	0	0	0	8
v_2	$\boxed{\frac{64}{10}}$	-8	$-\frac{32}{10}$	$\frac{12}{10}$	$\frac{54}{10}$	0
u_1	0	0	0	-1	0	0
u_2	0	0	0	0	-1	0
μ	$\frac{2}{10}$	0	$-\frac{1}{10}$	$\frac{1}{10}$	$\frac{2}{10}$	1
	$-\frac{2}{10}$	0	$\frac{1}{10}$	$-\frac{1}{10}$	$-\frac{2}{10}$	-1

18.2.3.2 Hildreth-d'Esopo Method

1. Principle

The strictly convex optimization problem

$$f(\underline{\mathbf{x}}) = \underline{\mathbf{x}}^{\mathrm{T}}\mathbf{C}\underline{\mathbf{x}} + \underline{\mathbf{p}}^{\mathrm{T}}\underline{\mathbf{x}} = \min!\,, \quad \mathbf{A}\underline{\mathbf{x}} \le \underline{\mathbf{b}} \tag{18.63}$$

has the dual problem (see 1., p. 926)

$$\psi(\underline{\mathbf{u}}) = \underline{\mathbf{u}}^{\mathrm{T}}\mathbf{E}\underline{\mathbf{u}} + \underline{\mathbf{h}}^{\mathrm{T}}\underline{\mathbf{u}} = \min! \quad \underline{\mathbf{u}} \ge 0 \quad \text{with} \tag{18.64a}$$

$$\mathbf{E} = \frac{1}{4}\mathbf{A}\mathbf{C}^{-1}\mathbf{A}^{\mathrm{T}}, \quad \underline{\mathbf{h}} = \frac{1}{2}\mathbf{A}\mathbf{C}^{-1}\underline{\mathbf{p}} + \underline{\mathbf{b}}. \tag{18.64b}$$

Matrix $\mathbf{E}$ is positive definite and it has positive diagonal elements $e_{ii} > 0$, $(i = 1, 2, \ldots, m)$. The variables $\underline{\mathbf{x}}$ and $\underline{\mathbf{u}}$ satisfy the following relation:

$$\underline{\mathbf{x}} = -\frac{1}{2}\mathbf{C}^{-1}(\mathbf{A}^{\mathrm{T}}\underline{\mathbf{u}} + \underline{\mathbf{p}}). \tag{18.65}$$

2. Solution by Iteration

The dual problem (18.64a), which contains only the condition $\underline{\mathbf{u}} \ge \underline{\mathbf{0}}$, can be solved by the following simple iteration method:

a) Substitute $\underline{\mathbf{u}}^1 \ge \underline{\mathbf{0}}$, (e.g., $\underline{\mathbf{u}}^1 = \underline{\mathbf{0}}$), $k = 1$.

b) Calculate u_i^{k+1} for $i = 1, 2, \ldots, m$ according to

$$w_i^{k+1} = -\frac{1}{e_{ii}}\left(\sum_{j=1}^{i-1} e_{ij}u_j^{k+1} + \frac{h_i}{2} + \sum_{j=i+1}^{m} e_{ij}u_j^k\right), \tag{18.66a}$$

$$u_i^{k+1} = \max\left\{0, w_i^{k+1}\right\}. \tag{18.66b}$$

c) Repeat step **b)** with $k+1$ instead of k until a stopping rule is satisfied, e.g., $\left|\psi(\underline{\mathbf{u}}^{k+1}) - \psi(\underline{\mathbf{u}}^k)\right| < \varepsilon$, $\varepsilon > 0$.
Under the assumption that there is an $\underline{\mathbf{x}}$ such that $\mathbf{A}\underline{\mathbf{x}} < \underline{\mathbf{b}}$, the sequence $\{\psi(\underline{\mathbf{u}}^k)\}$ converges to the minimum value $\psi_{\min}$ and sequence $\{\underline{\mathbf{x}}^k\}$ given by (18.65) converges to the solution $\underline{\mathbf{x}}^*$ of the original problem. The sequence $\{\underline{\mathbf{u}}^k\}$ is not always convergent.

18.2.4 Numerical Search Procedures

By using non-linear optimization procedures acceptable approximate solutions can be found with reasonable computing costs for several types of optimization problems. They are based on the principle of comparison of function values.

18.2.4.1 One-Dimensional Search

Several optimization methods contain the subproblem of finding the minimum of a real function $f(x)$ for $x \in [a,b]$. It is often sufficient to find an approximation $\overline{x}$ of the minimum point x^*.

1. Formulation of the Problem

A function $f(x)$, $x \in \mathbb{R}$, is called unimodal in $[a,b]$ if it has exactly one local minimum point on every closed subinterval $J \subseteq [a,b]$. Let f be a unimodal function on $[a,b]$ and x^* the global minimum point. Then an interval $[c,d] \subseteq [a,b]$ should be found with $x^* \in [c,d]$ such that $d - c < \varepsilon$, $\varepsilon > 0$.

2. Uniform Search

A positive integer n is chosen such that $\delta = \dfrac{b-a}{n+1} < \dfrac{\varepsilon}{2}$, and the values $f(x^k)$ for $x^k = a + k\delta$ $(k = 1,\dots,n)$ are calculated. If $f(x)$ is the smallest value among these function values, then the minimum point x^* is in the interval $[x-\delta, x+\delta]$. The number of required function values for the given accuracy can be estimated by

$$n > \frac{2(b-a)}{\varepsilon} - 1\,. \tag{18.67}$$

3. Golden Section Method, Fibonacci Method

The interval $[a,b] = [a_1,b_1]$ will be reduced step by step so that the new subinterval always contains the minimum point x^*. The points λ_1, μ_1 are determined in the interval $[a_1,b_1]$ as

$$\lambda_1 = a_1 + (1-\tau)(b_1 - a_1), \quad \mu_1 = a_1 + \tau(b_1 - a_1) \quad \text{with} \tag{18.68a}$$

$$\tau = \frac{1}{2}(\sqrt{5} - 1) \approx 0.618. \tag{18.68b}$$

This corresponds to the golden section. Two cases are distinguished:

a) If $f(\lambda_1) < f(\mu_1)$, then $a_2 = a_1$, $b_2 = \mu_1$ and $\mu_2 = \lambda_1$ are substituted. (18.69a)

b) If $f(\lambda_1) \ge f(\mu_1)$, then $a_2 = \lambda_1$, $b_2 = b_1$ and $\lambda_2 = \mu_1$ are substituted. (18.69b)

If $b_2 - a_2 \ge \varepsilon$, then the procedure is repeated with the interval $[a_2,b_2]$, where one value is already known, $f(\lambda_2)$ in case **a)** and $f(\mu_2)$ in case **b)**, from the first step. To determine an interval $[a_n,b_n]$, which contains the minimum point x^*, altogether n function values are calculated. From the requirement

$$\varepsilon > b_n - a_n = \tau^{n-1}(b_1 - a_1) \tag{18.70}$$

the necessary number of steps n can be estimated.
By using the golden section method, at most one more function value should be determined compared to the Fibonacci method. Instead of subdividing the interval according to the golden section, the interval is subdivided according to the *Fibonacci numbers* (see 5.4.1.5, p. 375, and 17.3.2.4, **4.**, p. 908).

18.2.4.2 Minimum Search in n-Dimensional Euclidean Vector Space

The search for an approximation of the minimum point $\underline{\mathbf{x}}^*$ of the problem $f(\underline{\mathbf{x}}) = \min!$, $\underline{\mathbf{x}} \in \mathbb{R}^n$, can be reduced to the solution of a sequence of one-dimensional optimization problems.
One takes

a) $\underline{\mathbf{x}} = \underline{\mathbf{x}}^1$, $\quad k = 1$, where $\underline{\mathbf{x}}^1$ is an appropriate initial approximation of $\underline{\mathbf{x}}^*$. (18.71a)

b) The one-dimensional problems

$$\varphi(\alpha_r) = f(x_1^{k+1}, \ldots, x_{r-1}^{k+1}, x_r^k + \alpha_r, x_{r+1}^k, \ldots, x_n^k) = \min! \text{ with } \quad \alpha_r \in \mathbb{R} \tag{18.71b}$$

are solved for $r = 1, 2, \ldots, n$. If $\bar{\alpha}_r$ is an exact or approximating minimum point of the r-th problem, then $x_r^{k+1} = x_r^k + \bar{\alpha}_r$ are substituted.

c) If two consecutive approximations are close enough to each other, i.e., with some vector norm,

$$||\underline{\mathbf{x}}^{k+1} - \underline{\mathbf{x}}^k|| < \varepsilon_1 \quad \text{or} \quad |f(\underline{\mathbf{x}}^{k+1}) - f(\underline{\mathbf{x}}^k)| < \varepsilon_2, \tag{18.71c}$$

then $\underline{\mathbf{x}}^{k+1}$ is an approximation of $\underline{\mathbf{x}}^*$. Otherwise step b) is repeated with $k+1$ instead of k. The one-dimensional problem in b) can be solved, by using the methods given in 18.2.4.1, p. 930.

18.2.5 Methods for Unconstrained Problems

The general optimization problem

$$f(\underline{\mathbf{x}}) = \min! \quad \text{for} \quad \underline{\mathbf{x}} \in \mathbb{R}^n \tag{18.72}$$

is considered with a continuously differentiable function f. Each method described in this section constructs, in general, an infinite sequence of points $\{\underline{\mathbf{x}}^k\} \in \mathbb{R}^n$, whose accumulation point is a stationary point. The sequence of points will be determined starting with a point $\underline{\mathbf{x}}^1 \in \mathbb{R}^n$ and according to the formula

$$\underline{\mathbf{x}}^{k+1} = \underline{\mathbf{x}}^k + \alpha_k \underline{\mathbf{d}}^k \quad (k = 1, 2, \ldots), \tag{18.73}$$

i.e., first a direction $\underline{\mathbf{d}}^k \in \mathbb{R}^n$ is determined at $\underline{\mathbf{x}}^k$ and the *step size* $\alpha_k \in \mathbb{R}$ indicates how far $\underline{\mathbf{x}}^{k+1}$ is from $\underline{\mathbf{x}}^k$ in the direction $\underline{\mathbf{d}}^k$. Such a method is called a *descent method*, if

$$f(\underline{\mathbf{x}}^{k+1}) < f(\underline{\mathbf{x}}^k) \quad (k = 1, 2, \ldots). \tag{18.74}$$

The equality $\nabla f(\underline{\mathbf{x}}) = 0$, where ∇ is the nabla operator (see 13.2.6.1, p. 715), characterizes a stationary point and can be used as a stopping rule for the iteration method.

18.2.5.1 Method of Steepest Descent

Starting from an actual point $\underline{\mathbf{x}}^k$, the direction $\underline{\mathbf{d}}^k$ in which the function has its steepest descent is

$$\underline{\mathbf{d}}^k = -\nabla f(\underline{\mathbf{x}}^k) \quad (18.75a) \qquad \text{and consequently} \quad \underline{\mathbf{x}}^{k+1} = \underline{\mathbf{x}}^k - \alpha_k \nabla f(\underline{\mathbf{x}}^k). \tag{18.75b}$$

A schematic representation of the *steepest descent method* with level lines $f(\underline{\mathbf{x}}) = f(\underline{\mathbf{x}}^i)$ is shown in **Fig. 18.6**.

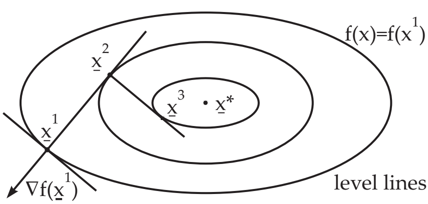


Figure 18.6

The step size α_k is determined by a line search, i.e., α_k is the solution of the one-dimensional problem:

$$f(\underline{\mathbf{x}}^k + \alpha \underline{\mathbf{d}}^k) = \min!, \quad \alpha \geq 0. \tag{18.76}$$

This problem can be solved by the methods given in 18.2.4, p. 930.

The steepest descent method (18.75b) converges relatively slowly. For every accumulation point $\underline{\mathbf{x}}^*$ of the sequence $\{\underline{\mathbf{x}}^k\}$, $\nabla f(\underline{\mathbf{x}}^*) = 0$. In the case of a quadratic objective function, i.e., $f(\underline{\mathbf{x}}) = \underline{\mathbf{x}}^{\mathrm{T}}\mathbf{C}\underline{\mathbf{x}} + \underline{\mathbf{p}}^{\mathrm{T}}\underline{\mathbf{x}}$, the method has the special form:

$$\underline{\mathbf{x}}^{k+1} = \underline{\mathbf{x}}^k + \alpha_k \underline{\mathbf{d}}^k \quad (18.77a) \quad \text{with } \underline{\mathbf{d}}^k = -(2\mathbf{C}\underline{\mathbf{x}}^k + \underline{\mathbf{p}}) \text{ and } \alpha_k = \frac{\underline{\mathbf{d}}^{k\,\mathrm{T}}\underline{\mathbf{d}}^k}{2\underline{\mathbf{d}}^{k\,\mathrm{T}}\mathbf{C}\underline{\mathbf{d}}^k}. \tag{18.77b}$$

18.2.5.2 Application of the Newton Method

Suppose that at the actual approximation point $\underline{\mathbf{x}}^k$ the function f is approximated by a quadratic function:

$$q(\underline{\mathbf{x}}) = f(\underline{\mathbf{x}}^k) + (\underline{\mathbf{x}} - \underline{\mathbf{x}}^k)^{\mathrm{T}} \nabla f(\underline{\mathbf{x}}^k) + \frac{1}{2}(\underline{\mathbf{x}} - \underline{\mathbf{x}}^k)^{\mathrm{T}} \mathbf{H}(\underline{\mathbf{x}}^k)(\underline{\mathbf{x}} - \underline{\mathbf{x}}^k). \tag{18.78}$$

Here $\mathbf{H}(\underline{\mathbf{x}}^k)$ is the Hessian matrix, i.e., the matrix of second partial derivatives of f at the point $\underline{\mathbf{x}}^k$. If $\mathbf{H}(\underline{\mathbf{x}}^k)$ is positive definite, then $q(\underline{\mathbf{x}})$ has an absolute minimum at $\underline{\mathbf{x}}^{k+1}$ with $\nabla q(\underline{\mathbf{x}}^{k+1}) = 0$, therefore one gets the Newton method:

$$\underline{\mathbf{x}}^{k+1} = \underline{\mathbf{x}}^k - \mathbf{H}^{-1}(\underline{\mathbf{x}}^k)\nabla f(\underline{\mathbf{x}}^k) \quad (k = 1, 2, \ldots), \quad \text{i.e.,} \tag{18.79a}$$

$$\underline{\mathbf{d}}^k = -\mathbf{H}^{-1}(\underline{\mathbf{x}}^k)\nabla f(\underline{\mathbf{x}}^k) \quad \text{and} \quad \alpha_k \text{ in (18.73).} \tag{18.79b}$$

The Newton method converges fast but it has the following disadvantages:
a) The matrix $\mathbf{H}(\underline{\mathbf{x}}^k)$ must be positive definite.
b) The method converges only for sufficiently good initial points.
c) The step size can not influenced.
d) The method is not a descent method.
e) The computational cost of computing the inverse of $\mathbf{H}^{-1}(\underline{\mathbf{x}}^k)$ is fairly high.
Some of these disadvantages can be reduced by the following version of the *damped Newton method* (see also 19.2.2.2, p. 962):

$$\underline{\mathbf{x}}^{k+1} = \underline{\mathbf{x}}^k - \alpha_k \mathbf{H}^{-1}(\underline{\mathbf{x}}^k)\nabla f(\underline{\mathbf{x}}^k) \quad (k = 1, 2, \ldots). \tag{18.80}$$

The relaxation factor α_k can be determined, for example, by the principle given earlier (see 18.2.5.1, p. 931).

18.2.5.3 Conjugate Gradient Methods

Two vectors $\underline{\mathbf{d}}^1, \underline{\mathbf{d}}^2 \in \mathbb{R}^n$ are called *conjugate vectors* with respect to a symmetric, positive definite matrix $\mathbf{C}$, if

$$\underline{\mathbf{d}}^{1\mathrm{T}} \mathbf{C} \underline{\mathbf{d}}^2 = 0. \tag{18.81}$$

If $\underline{\mathbf{d}}^1, \underline{\mathbf{d}}^2, \ldots, \underline{\mathbf{d}}^n$ are pairwise conjugate vectors with respect to a matrix $\mathbf{C}$, then the convex quadratic problem $q(\underline{\mathbf{x}}) = \underline{\mathbf{x}}^{\mathrm{T}}\mathbf{C}\underline{\mathbf{x}} + \underline{\mathbf{p}}^{\mathrm{T}}\underline{\mathbf{x}}$, $\underline{\mathbf{x}} \in \mathbb{R}^n$, can be solved in n steps if a sequence $\underline{\mathbf{x}}^{k+1} = \underline{\mathbf{x}}^k + \alpha_k \underline{\mathbf{d}}^k$ starting from $\underline{\mathbf{x}}^1$ is constructed, where α_k is the optimal step size. Under the assumption that $f(\underline{\mathbf{x}})$ is approximately quadratic in the neighborhood of $\underline{\mathbf{x}}^*$, i.e., $\mathbf{C} \approx \frac{1}{2}\mathbf{H}(\underline{\mathbf{x}}^*)$, the method developed for quadratic objective functions can also be applied for more general functions $f(\underline{\mathbf{x}})$, without the explicit use of the matrix $\mathbf{H}(\underline{\mathbf{x}}^*)$.
The conjugate gradient method has the following steps:

a) $$\underline{\mathbf{x}}^1 \in \mathbb{R}^n, \quad \underline{\mathbf{d}}^1 = -\nabla f(\underline{\mathbf{x}}^1), \tag{18.82}$$

where $\underline{\mathbf{x}}^1$ is an appropriate initial approximation for $\underline{\mathbf{x}}^*$.

b) $$\underline{\mathbf{x}}^{k+1} = \underline{\mathbf{x}}^k + \alpha_k \underline{\mathbf{d}}^k \;\; (k = 1, \ldots, n) \text{ with } \alpha_k \geq 0 \text{ so that } f(\underline{\mathbf{x}}^k + \alpha \underline{\mathbf{d}}^k) \text{ will be minimized.} \tag{18.83a}$$

$$\underline{\mathbf{d}}^{k+1} = -\nabla f(\underline{\mathbf{x}}^{k+1}) + \mu_k \underline{\mathbf{d}}^k \;\; (k = 1, \ldots, n-1) \text{ with} \tag{18.83b}$$

$$\mu_k = \frac{\nabla f(\underline{\mathbf{x}}^{k+1})^{\mathrm{T}} \nabla f(\underline{\mathbf{x}}^{k+1})}{\nabla f(\underline{\mathbf{x}}^k)^{\mathrm{T}} \nabla f(\underline{\mathbf{x}}^k)} \quad \text{and} \quad \underline{\mathbf{d}}^{n+1} = -\nabla f(\underline{\mathbf{x}}^{n+1}). \tag{18.83c}$$

c) Repeating steps **b)** with $\underline{\mathbf{x}}^{n+1}$ and $\underline{\mathbf{d}}^{n+1}$ instead of $\underline{\mathbf{x}}^1$ and $\underline{\mathbf{d}}^1$.

18.2.5.4 Method of Davidon, Fletcher and Powell (DFP)

With the DFP method, a sequence of points starting from $\underline{\mathbf{x}}^1 \in \mathbb{R}^n$ is determined according to the formula

$$\underline{\mathbf{x}}^{k+1} = \underline{\mathbf{x}}^k - \alpha_k \mathbf{M}_k \nabla f(\underline{\mathbf{x}}^k) \quad (k = 1, 2, \ldots). \tag{18.84}$$

Here, $\mathbf{M}_k$ is a symmetric, positive definite matrix. The idea of the method is a stepwise approximation of the inverse Hessian matrix by matrices $\mathbf{M}_k$ in the case when $f(\underline{\mathbf{x}})$ is a quadratic function. Starting

with a symmetric, positive definite matrix $\mathbf{M}_1$, e.g., $\mathbf{M}_1 = \mathbf{I}$ ($\mathbf{I}$ is the unit matrix), the matrix $\mathbf{M}_k$ is determined from $\mathbf{M}_{k-1}$ by adding a correction matrix of rank two

$$\mathbf{M}_k = \mathbf{M}_{k-1} + \frac{\underline{\mathbf{v}}^k \underline{\mathbf{v}}^{k\,\mathrm{T}}}{\underline{\mathbf{v}}^{k\,\mathrm{T}} \underline{\mathbf{v}}^k} - \frac{(\mathbf{M}_{k-1}\underline{\mathbf{w}}^k)(\mathbf{M}_{k-1}\underline{\mathbf{w}}^k)^{\mathrm{T}}}{\underline{\mathbf{w}}^{k\,\mathrm{T}} \mathbf{M}_k \underline{\mathbf{w}}^k} \qquad (18.85)$$

with $\underline{\mathbf{v}}^k = \underline{\mathbf{x}}^k - \underline{\mathbf{x}}^{k-1}$ and $\underline{\mathbf{w}}^k = \nabla f(\underline{\mathbf{x}}^k) - \nabla f(\underline{\mathbf{x}}^{k-1})$ $(k = 2, 3, \ldots)$. The step size α_k is obtained from

$$f(\underline{\mathbf{x}}^k - \alpha \mathbf{M}_k \nabla f(\underline{\mathbf{x}}^k)) = \min!\,, \quad \alpha \geq 0. \qquad (18.86)$$

If $f(\underline{\mathbf{x}})$ is a quadratic function, then the DFP method becomes the conjugate gradient method with $\mathbf{M}_1 = \mathbf{I}$.

18.2.6 Evolution Strategies

18.2.6.1 Evolution Principles

Evolution strategies are examples of stochastic optimization processes imitating natural evolution. They are based on the principles of mutation, recombination and selection.

1. Mutation

From a parent point $\underline{\mathbf{x}}_P$ a offspring (descendant) $\underline{\mathbf{x}}_O = \underline{\mathbf{x}}_P + \underline{\mathbf{d}}$ is formed by applying a random variation $\underline{\mathbf{d}}$. The components of $\underline{\mathbf{d}}$ are $(0, \sigma_i^2)$ normally distributed random variables $Z(0, \sigma_i^2)$ determined newly at every mutation:

$$\underline{\mathbf{d}} = \begin{pmatrix} d_1 \\ d_2 \\ \vdots \\ d_n \end{pmatrix} = \begin{pmatrix} Z(0, \sigma_1^2) \\ Z(0, \sigma_2^2) \\ \vdots \\ Z(0, \sigma_n^2) \end{pmatrix} = \begin{pmatrix} Z(0,1) \cdot \sigma_1 \\ Z(0,1) \cdot \sigma_2 \\ \vdots \\ Z(0,1) \cdot \sigma_n \end{pmatrix}. \qquad (18.87)$$

With a normally distributed $\underline{\mathbf{d}}$ small changes have high probabilities while large changes occur very rarely. The changes are controlled by the standard deviation σ_i.

2. Recombination

From the population of μ parents offspring can be obtained by mixing the information from two or more parents, which are randomly selected. The recombination can follow two types of changes.
At *intermediate recombination* a offspring becomes as weighted average of ϱ randomly chosen parents:

$$\underline{\mathbf{x}}_O = \sum_{i=1}^{\varrho} \alpha_i \, \underline{\mathbf{x}}_{P_i}, \quad \sum_{i=1}^{\varrho} \alpha_i = 1, \quad 2 \leq \varrho \leq \mu\,. \qquad (18.88)$$

At a *discrete* recombination of ϱ parents the i-th component of a offspring $\underline{\mathbf{x}}_O$ is determined by the i-th component of a randomly chosen parent:

$$x_{iO} = x_{iP_j}, \quad j \in \{1, \ldots, \varrho\}, \quad i = 1, \ldots, n. \qquad (18.89)$$

3. Selection

By using mutation and recombination, a set of offspring is formed randomly. In a subsequent selection process the *objective function* $f(\underline{\mathbf{x}})$ serves as a measure to compare the fitness of the individuals. The fittest individuals are selected for the next generation. At certain strategies only the offspring take part in the selection. Other strategies consider also the parents (see also [18.12]).

18.2.6.2 Evolution Algorithms

Every evolution strategy is based on the following algorithm:

a) Determination of an appropriate starting population consisting of μ individuals. These are the first generation of parents. $X_P^1 = \{\underline{\mathbf{x}}_{P_1}^1, \ldots, \underline{\mathbf{x}}_{P_\mu}^1\}$.

b) In the k-th step the creation of λ offspring $X_O^k = \{\underline{\mathbf{x}}_{O_1}^k, \ldots, \underline{\mathbf{x}}_{O_\lambda}^k\}$ by mutation and recombination of parents of the actual generation $X_P^k = \{\underline{\mathbf{x}}_{P_1}^k, \ldots, \underline{\mathbf{x}}_{P_\lambda}^k\}$.

c) Application of selection to get the best μ individuals for the next parent generation $X_P^{k+1} = \{\underline{\mathbf{x}}_{P_1}^{k+1}, \dots, \underline{\mathbf{x}}_{P_\mu}^{k+1}\}$.

d) Repeating steps b) and c) until stopping rule is satisfied. It can be fulfilling an optimal criterium of the optimization problem, or to reach a given number of generations, or exceeding a given computer time, etc.

18.2.6.3 Classification of Evolution Strategies

Every evolution strategy is characterized by a sequence of parameters. Essential parameters are the size of the population μ, the number of the offspring λ, the number of parents ϱ taking part in recombination and rules of making mutation, recombination and selection. To distinguish different types of strategies a special notation is commonly used. For the strategies using only mutation in producing offspring the $(\mu+\lambda)$, or (μ,λ) strategy notation is used. Strategies $(\mu+\lambda)$ and (μ,λ) differ from each other in the type of selection. At strategy (μ,λ) the selection of the new generation is made only among the offspring, while at strategy $(\mu+\lambda)$ the parents are also involved.
For strategies using recombination the number ϱ of the parents, which are involved, is seen in the notation $(\mu/\varrho+\lambda)$- and $(\mu/\varrho,\lambda)$-strategy.

18.2.6.4 Generating Random Numbers

For the numerical evaluation of evolution procedures *uniformly* and *normally* distributed random variables are needed. Values of uniformly distributed variables can be got by the methods given in subchapter 16.3.5.2, p. 843. Normally distributed random variables can be produced from uniform variables in the following way:
Box-Muller Method: If G_1 and G_2 are uniformly distributed random numbers in the interval $[0,1]$, then the following two equations give two statistically independent normally distributed $(0,\sigma^2)$ random numbers $Z_1(0,\sigma^2)$ and $Z_2(0,\sigma^2)$:

$$Z_1(0,\sigma^2) = \sigma\sqrt{-2\ln G_1}\cos(2\pi G_2) \quad \text{and} \quad Z_2(0,\sigma^2) = \sigma\sqrt{-2\ln G_1}\sin(2\pi G_2)\,. \tag{18.90}$$

18.2.6.5 Application of Evolution Strategies

In the practice optimization problems have usually high complexity. Here the conventional optimization processes described in 18.2.5, p. 931 are often not appropriate. Evolution strategies belong to the differentiation-free solution methods, which are based on comparisons of the values of the objective function. They have simple conditions on the structure of the objective function. The objective function does not need to be differentiable or continuous. So the evolution strategies are appropriate for a wide spectrum of optimization problems.

The application of evolution strategies is not restricted to unconstrained continuous optimization problems. Optimization problems with constraints can also be handled, where the constraints are enforced by penalty terms in the objective function (see Penalty and Barrier Methods in 18.2.8, p. 940).
Another field of application is the discrete optimization, where some or all components of $\underline{\mathbf{x}}$ can take their values from a discrete set. One possible mutation mechanism is to replace the value of a discrete component by one of its neighboring values with the same probability.

18.2.6.6 (1 + 1)-Mutation-Selection Strategy

This method is similar to the gradient method discussed in 18.2.5, p. 931 with the difference that the direction $\underline{\mathbf{d}}^k$ is a normally distributed random vector. The population consists of a single individual which produces one offspring at every generation.

1. Mutation Step

In generation k a offspring is obtained from a parent by adding a normally distributed random vector:

$$\underline{\mathbf{x}}_O^k = \underline{\mathbf{x}}_P^k + \alpha\underline{\mathbf{d}}^k\,. \tag{18.91}$$

The factor α is a parameter by which the speed of the convergence can be affected. α is considered as the step size of the mutation.

2. Selection Step

The new parent of the next generation $(k+1)$ is selected by comparing the objective function values of both individuals, i.e., from the parent with the formula:

$$\underline{\mathbf{x}}_P^{k+1} = \begin{cases} \underline{\mathbf{x}}_O^k & \text{if } f(\underline{\mathbf{x}}_O^k) < f(\underline{\mathbf{x}}_P^k)\,, \\ \underline{\mathbf{x}}_P^k & \text{otherwise}\,. \end{cases} \qquad (18.92)$$

The procedure stops if no better offspring arrives over a given number of generations. The step size α can be increased if the mutation results mostly in improved offspring. At small improvements the value of α should be decreased.

3. Step Size Control

The choice of the mutation step size α is of important influence to the convergence properties of the evolution method. While large step sizes are recommended in order to have fast convergence, small step size is required in the close neighborhood of the optimum or in regions of fast changing or oscillating of the objective function. The optimal step size depends on the problem. Too small steps lead to stagnation, too large steps may result in overshooting of the evolution process.

1. 1/5-Success rule: The ratio of the number of successful mutations and the total number of mutations in the last step defines the rate of success q. If $q > 1/5$, then the step size can be increased. For smaller success rate, α should be decreased:

$$\alpha_{k+1} = \begin{cases} c \cdot \alpha_k, & q < \dfrac{1}{5}\,, \\ \dfrac{1}{c} \cdot \alpha_k, & q > \dfrac{1}{5} \end{cases} \quad \text{with} \quad c = 0,8 \ldots 0,85. \qquad (18.93)$$

2. Mutative Step Size Determination: The rule of 1/5 is a rough choice, and considering any concrete problem it is not always satisfactory. In an extended model the step size α and the standard deviations σ_i, $i = 1, 2 \ldots, n$ are in correlation. Here α and σ_i are multiplied with equal probability by one of the factors c, 1, $1/c$, where $c = 1,1 \ldots 1,5$. Further information see [18.12].

18.2.6.7 Population Strategies

The (1+1) strategy presented in the preceding paragraph reflects the principles of the natural evolution only in a very simplified form. With the extension to population models further properties of the evolution process can be considered. A large number of individuals in an evolution process assures that different regions of the solution space will be searched.

1. $(\mu + \lambda)$-Evolution Strategy

The $(\mu + \lambda)$ strategy is a generalization of the $(1 + 1)$ strategy. From the μ parents of the current generation $X_P^k = \{\underline{\mathbf{x}}_{P_1}^k, \ldots, \underline{\mathbf{x}}_{P_\mu}^k\}$ a set of λ parents is chosen randomly with equal probability. Repeated choices are allowed and even necessary in the case if $\mu < \lambda$. By mutation, λ offspring $X_O^k = \{\underline{\mathbf{x}}_{O_1}^k, \ldots, \underline{\mathbf{x}}_{O_\lambda}^k\}$ are produced. From the selection set $X_O^k \cup X_P^k$ the best μ individuals are chosen to take over into the next generation.

Since the parents are also taken in the selection, the quality of the population from a generation to the next one cannot be worse. The $(\mu + \lambda)$ strategy has the property that it keeps an already found local optimum, since large mutation steps, which are required to leave the optimal point, have very small probability. It means, that an individual can have an infinite life. This behavior can be avoided by adding penalty terms to the objective function values of parents that increase from generation to generation. In this way the aging of individuals can be simulated.

2. $(\mu\,, \lambda)$-Evolution Strategy

In contrary to the $(\mu + \lambda)$ strategy the selection step is now made among the λ offspring, to choose the μ individuals for the next generation, i.e., in this strategy the parents do not survive. Therefore $\lambda > \mu$

must hold. The values of the objective function for the offspring can be larger than that for the parents. This procedure can depart from local optima.

Selection Pressure: The ratio of the individuals taking part in the selection and the size of the population defines the selection pressure S:

$$S = \begin{cases} \dfrac{\mu+\lambda}{\mu} & \text{for } (\mu+\lambda)\text{-strategy}\,, \\ \dfrac{\lambda}{\mu} & \text{for } (\mu\,,\ \lambda)\text{-strategy} \end{cases} \quad \text{with} \quad 1 \le S < \infty. \tag{18.94}$$

If the selection pressure is close to 1, then the selection step has almost no impact. A large number of offspring $\lambda \gg \mu$ results in a strong selection pressure, since from the set of the present individuals only few will survive into the next generation.

3. $(\mu/\varrho + \lambda)$- and $(\mu/\varrho\,,\ \lambda)$-Evolution Strategies with Recombination

With the concept of recombination some relations are built between the individuals of a population, so the information of several parents are mixed in a offspring.

In order to get a offspring, ϱ parents are chosen from the set of parents X_P^k with the same probability. It is assumed that for every member of the λ offspring a separate choice of ϱ parents is taken. The offspring is a discrete or an intermediate recombination of the chosen parents. The offspring produced in this way will be mutated and enters the selection process.

In the previously described $(\mu + \lambda)$ and (μ, λ) strategies each individual is the result of a series of mutations applied to one individual of the first generation of parents. So, a wider evolution step is possible only through many generations. Wider evolution steps are possible with recombination. Especially, when the parents have large distances among each other, the offspring are formed with completely new properties.

4. Evolution Strategies with more Populations

The principles of evolution can be expanded formally to the dimension of populations. Instead of individuals now populations are in competition. So, the evolution process has two steps. It is shown in the expanded notation: $[\mu_2/\varrho_2 \overset{+}{,} \lambda_2(\mu_1/\varrho_1 \overset{+}{,} \lambda_1)]$.

From a set of μ_2 parent populations a set of λ_2 offspring populations is created by recombination of ϱ_2 populations that are chosen randomly for each offspring population. In these λ_2 offspring populations the optimization is performed using a $(\mu_1/\varrho_1 + \lambda_1)$ or $(\mu_1/\varrho_1\,,\ \lambda_1)$ strategy. After a given number of generations the best populations are chosen based on an appropriate criterium. The comparison of populations can be done by considering the values of the objective function of the best individual or by the population mean.

18.2.7 Gradient Method for Problems with Inequality Type Constraints

If the problem

$$f(\underline{\mathbf{x}}) = \min! \quad \text{subject to the constraints} \quad g_i(\underline{\mathbf{x}}) \le 0 \quad (i = 1, \ldots, m) \tag{18.95}$$

has to be solved by an iteration method of the type

$$\underline{\mathbf{x}}^{k+1} = \underline{\mathbf{x}}^k + \alpha_k \underline{\mathbf{d}}^k \quad (k = 1, 2, \ldots) \tag{18.96}$$

then two additional rules must be considered because of the bounded feasible set:

1. The direction $\underline{\mathbf{d}}^k$ must be a feasible descent direction at $\underline{\mathbf{x}}^k$.

2. The step size α_k must be determined so that $\underline{\mathbf{x}}^{k+1}$ is in M.

The different methods based on the formula (18.96) differ from each other only in the construction of the direction $\underline{\mathbf{d}}^k$. To ensure the feasibility of the sequence $\{\underline{\mathbf{x}}^k\} \subset M$, α_k' and α_k'' are determined in the following way:

$$\alpha_k' \quad \text{from} \quad f(\underline{\mathbf{x}}^k + \alpha \underline{\mathbf{d}}^k) = \min!\,, \quad \alpha \ge 0$$

$$\alpha_k'' = \max\{\alpha \in \mathbb{R} : \underline{\mathbf{x}}^k + \alpha\underline{\mathbf{d}}^k \in M\}. \tag{18.97}$$

Then

$$\alpha_k = \min\{\alpha_k', \alpha_k''\}. \tag{18.98}$$

If there is no feasible descent direction $\underline{\mathbf{d}}^k$ in a certain step k, then $\underline{\mathbf{x}}^k$ is a stationary point.

18.2.7.1 Method of Feasible Directions

1. Direction Search Program

A feasible descent direction $\underline{\mathbf{d}}^k$ at point $\underline{\mathbf{x}}^k$ can be determined by the solution of the following optimization problem:

$$\sigma = \min!, \tag{18.99}$$

$$\nabla g_i(\underline{\mathbf{x}}^k)^{\mathrm{T}}\underline{\mathbf{d}} \le \sigma, \quad i \in I_0(\underline{\mathbf{x}}^k), \quad (18.100a) \qquad \nabla f(\underline{\mathbf{x}}^k)^{\mathrm{T}}\underline{\mathbf{d}} \le \sigma, \quad (18.100b) \qquad ||\underline{\mathbf{d}}|| \le 1. \quad (18.100c)$$

If $\sigma < 0$ for the result $\underline{\mathbf{d}} = \underline{\mathbf{d}}^k$ of this *direction search program*, then (18.100a) ensures feasibility and (18.100b) ensures the descending property of $\underline{\mathbf{d}}^k$. The feasible set for the direction search program is bounded by the normalizing condition (18.100c). If $\sigma = 0$, then $\underline{\mathbf{x}}^k$ is a stationary point, since there is no feasible descent direction at $\underline{\mathbf{x}}^k$.

A direction search program, defined by (18.100a,b,c), can result in a zig-zag behavior of the sequence $\underline{\mathbf{x}}^k$, which can be avoided if the index set $I_0(\underline{\mathbf{x}}^k)$ is replaced by the index set

$$I_{\varepsilon_k}(\underline{\mathbf{x}}^k) = \{i \in \{1, \ldots, m\} : \ -\varepsilon_k \le g_i(\underline{\mathbf{x}}^k) \le 0\}, \quad \varepsilon_k \ge 0 \tag{18.101}$$

which are the so-called ε_k active constraints in $\underline{\mathbf{x}}^k$. Thus, local directions of descent are excluded which are going from $\underline{\mathbf{x}}^k$ and lead closer to the boundaries of M consisting of the ε_k active constraints **(Fig. 18.7)**.

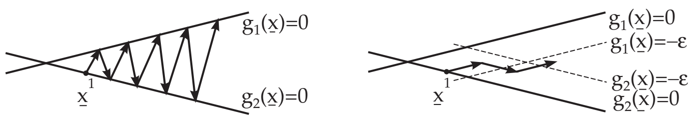


Figure 18.7

If $\sigma = 0$ is a solution of (18.100a,b,c) after these modifications, then $\underline{\mathbf{x}}^k$ is a stationary point only if $I_0(\underline{\mathbf{x}}^k) = I_{\varepsilon_k}(\underline{\mathbf{x}}^k)$. Otherwise ε_k must be decreased and the direction search program must be repeated.

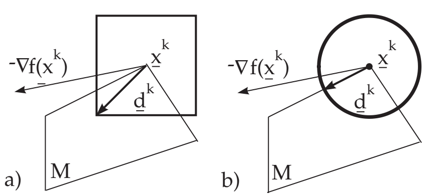


Figure 18.8

2. Special Case of Linear Constraints

If the functions $g_i(\underline{\mathbf{x}})$ are linear, i.e., $g_i(\underline{\mathbf{x}}) = \underline{\mathbf{a}}_i^{\mathrm{T}}\underline{\mathbf{x}} - b_i$, then a simpler direction search method can be established:

$$\sigma = \nabla f(\underline{\mathbf{x}}^k)^{\mathrm{T}}\underline{\mathbf{d}} = \min! \quad \text{with} \tag{18.102}$$

$$\underline{\mathbf{a}}_i^{\mathrm{T}}\underline{\mathbf{d}} \le 0, \quad i \in I_0(\underline{\mathbf{x}}^k) \ \text{ or } \ i \in I_{\varepsilon_k}(\underline{\mathbf{x}}^k), \tag{18.103a}$$

$$||\underline{\mathbf{d}}|| \le 1. \tag{18.103b}$$

The effect of the choice of different norms $||\underline{\mathbf{d}}|| = \max\{|d_i|\} \le 1$ or $||\underline{\mathbf{d}}|| = \sqrt{\underline{\mathbf{d}}^{\mathrm{T}}\underline{\mathbf{d}}} \le 1$ is shown in **Fig. 18.8a,b**.

In a certain sense, the best choice is the norm $||\underline{\mathbf{d}}|| = ||\underline{\mathbf{d}}||_2 = \sqrt{\underline{\mathbf{d}}^{\mathrm{T}}\underline{\mathbf{d}}}$, since by the direction search program the direction $\underline{\mathbf{d}}^k$ is obtained, which forms the smallest angle with $-\nabla f(\underline{\mathbf{x}}^k)$. In this case the

direction search program is not linear and requires higher computational costs. With the choice $||\underline{\mathbf{d}}|| = ||\underline{\mathbf{d}}||_\infty = \max\{|d_i|\} \le 1$ a system of linear constraints $-1 \le d_i \le 1$, $(i = 1, \dots, n)$ is obtained, so the direction search program can be solved, e.g., by the simplex method.
In order to ensure that the method of feasible directions for a quadratic optimization problem $f(\underline{\mathbf{x}}) = \underline{\mathbf{x}}^{\mathrm{T}}\mathbf{C}\underline{\mathbf{x}} + \underline{\mathbf{p}}^{\mathrm{T}}\underline{\mathbf{x}} = \min!$ with $\mathbf{A}\underline{\mathbf{x}} \le \underline{\mathbf{b}}$ results in a solution in finitely many steps, the direction search program is completed by the following conjugate requirements: If $\alpha_{k-1} = \alpha'_{k-1}$ holds in a step, i.e., $\underline{\mathbf{x}}^k$ is an "interior" point, then the condition

$$\underline{\mathbf{d}}^{k-1^{\mathrm{T}}}\mathbf{C}\underline{\mathbf{d}} = 0 \tag{18.104}$$

is added to the direction search program. Furthermore the corresponding conditions are kept from the previous steps. If in a later step $\alpha_k = \alpha''_k$ then the condition (18.104) is removed.

■ $f(\underline{\mathbf{x}}) = x_1^2 + 4x_2^2 - 10x_1 - 32x_2 = \min!$ $\quad g_1(\underline{\mathbf{x}}) = -x_1 \le 0,\ \ g_2(\underline{\mathbf{x}}) = -x_2 \le 0,$
$g_3(\underline{\mathbf{x}}) = x_1 + 2x_2 - 7 \le 0,\ \ g_4(\underline{\mathbf{x}}) = 2x_1 + x_2 - 8 \le 0.$

Step 1: Starting with $\underline{\mathbf{x}}^1 = (3,0)^{\mathrm{T}},\ \ \nabla f(\underline{\mathbf{x}}^1) = (-4,-32)^{\mathrm{T}},\ \ I_0(\underline{\mathbf{x}}^1) = \{2\}.$

Direction search program: $\left\{\begin{matrix} -4d_1 - 32d_2 = \min! \\ -d_2 \le 0,\ \ ||\underline{\mathbf{d}}||_\infty \le 1 \end{matrix}\right\} \Longrightarrow \underline{\mathbf{d}}^1 = (1,1)^{\mathrm{T}}.$

Minimizing constant: $\alpha'_k = -\dfrac{\underline{\mathbf{d}}^{k^{\mathrm{T}}}\nabla f(\underline{\mathbf{x}}^k)}{2\underline{\mathbf{d}}^{k^{\mathrm{T}}}\mathbf{C}\underline{\mathbf{d}}^k}$ with $\mathbf{C} = \begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix}.$

Maximal feasible step size: $\alpha''_k = \min\left\{\dfrac{-g_i(\underline{\mathbf{x}}^k)}{\underline{\mathbf{a}}_i^{\mathrm{T}}\underline{\mathbf{d}}^k}:\ \text{for } i \text{ such that } \underline{\mathbf{a}}_i^{\mathrm{T}}\underline{\mathbf{d}}^k > 0\right\}, \alpha'_1 = \dfrac{18}{5},\ \ \alpha''_1 = \dfrac{2}{3} \Longrightarrow$

$\alpha_1 = \min\left\{\dfrac{18}{5}, \dfrac{2}{3}\right\} = \dfrac{2}{3},\ \ \underline{\mathbf{x}}^2 = \left(\dfrac{11}{3}, \dfrac{2}{3}\right)^{\mathrm{T}}.$

Step 2: $\nabla f(\underline{\mathbf{x}}^2) = \left(-\dfrac{8}{3}, -\dfrac{80}{3}\right)^{\mathrm{T}},\quad I_0(\underline{\mathbf{x}}^2) = \{4\}.$

Direction search program: $\left\{\begin{matrix} -\dfrac{8}{3}d_1 - \dfrac{80}{3}d_2 = \min! \\ 2d_1 + d_2 \le 0,\ \ ||\underline{\mathbf{d}}||_\infty \le 1 \end{matrix}\right\} \Longrightarrow \underline{\mathbf{d}}^2 = \left(-\dfrac{1}{2}, 1\right)^{\mathrm{T}}, \alpha'_2 = \dfrac{152}{51},\ \ \alpha''_2 = \dfrac{4}{3} \Longrightarrow$

$\alpha_2 = \dfrac{4}{3},\ \ \underline{\mathbf{x}}^3 = (3,2)^{\mathrm{T}}.$

Step 3: $\nabla f(\underline{\mathbf{x}}^3) = (-4,-16)^{\mathrm{T}},\quad I_0 = (\underline{\mathbf{x}}^3) = \{3,4\}.$
Direction search program:
$\left\{\begin{matrix} -4d_1 - 16d_2 = \min! \\ d_1 + 2d_2 \le 0,\ \ 2d_1 + d_2 \le 0,\ \ ||\underline{\mathbf{d}}||_\infty \le 1 \end{matrix}\right\} \Longrightarrow \underline{\mathbf{d}}^3 =$
$\left(-1, \dfrac{1}{2}\right)^{\mathrm{T}}, \alpha'_3 = 1,\ \ \alpha''_3 = 3 \Longrightarrow \alpha^3 = 1,\ \ \underline{\mathbf{x}}^4 = \left(2, \dfrac{5}{2}\right)^{\mathrm{T}}.$
The next direction search program results in $\sigma = 0$. Here the minimum point is $\underline{\mathbf{x}}^* = \underline{\mathbf{x}}^4$ **(Fig. 18.9)**.

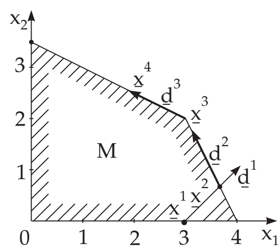


Figure 18.9

18.2.7.2 Gradient Projection Method

1. Formulation of the Problem and Solution Principle

Suppose the convex optimization problem

$$f(\underline{\mathbf{x}}) = \min! \quad \text{with} \quad \underline{\mathbf{a}}_i^{\mathrm{T}}\underline{\mathbf{x}} \le b_i, \tag{18.105}$$

for $i = 1, \dots, m$ is given. A feasible descent direction $\underline{\mathbf{d}}^k$ at the point $\underline{\mathbf{x}}^k \in M$ is determined in the following way:
If $-\nabla f(\underline{\mathbf{x}}^k)$ is a feasible direction, then $\underline{\mathbf{d}}^k = -\nabla f(\underline{\mathbf{x}}^k)$ is selected. Otherwise $\underline{\mathbf{x}}^k$ is on the boundary

of M and $-\nabla f(\underline{\mathbf{x}}^k)$ points outward from M. The vector $-\nabla f(\underline{\mathbf{x}}^k)$ is projected by a linear mapping $\mathbf{P}_k$ into a linear submanifold of the boundary of M defined by a subset of active constraints of $\underline{\mathbf{x}}^k$. **Fig. 18.10a** shows a projection into an edge, **Fig. 18.10b** shows a projection into a face. Supposing non-degeneracy, i.e., if the vectors $\underline{\mathbf{a}}_i$, $i \in I_0(\underline{\mathbf{x}})$ are linearly independent for every $\underline{\mathbf{x}} \in \mathbb{R}^n$, such a projection is given by

$$\underline{\mathbf{d}}^k = -\mathbf{P}_k \nabla f(\underline{\mathbf{x}}^k) = -\left(\mathbf{I} - {\mathbf{A}_k}^{\mathrm{T}}(\mathbf{A}_k{\mathbf{A}_k}^{\mathrm{T}})^{-1}\mathbf{A}_k\right)\nabla f(\underline{\mathbf{x}}^k). \tag{18.106}$$

Here, $\mathbf{A}_k$ consists of all vectors ${\underline{\mathbf{a}}_i}^{\mathrm{T}}$, whose corresponding constraints form the sub-manifold, into which $-\nabla f(\underline{\mathbf{x}}_k)$ should be projected.

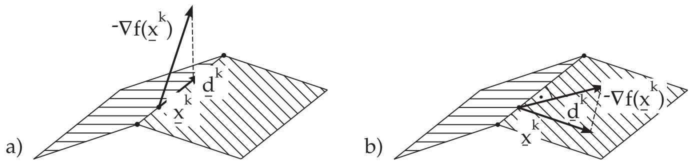


Figure 18.10

2. Algorithm

The gradient projection method consists of the following steps, starting with $\underline{\mathbf{x}}^1 \in M$ and substituting $k = 1$ and proceeding in accordance to the following scheme:

I: If $-\nabla f(\underline{\mathbf{x}}^k)$ is a feasible direction, then $\underline{\mathbf{d}}^k = -\nabla f(\underline{\mathbf{x}}^k)$ is substituted, and continued with **III**. Otherwise $\mathbf{A}_k$ is constructed from the vectors ${\underline{\mathbf{a}}_i}^{\mathrm{T}}$ with $i \in I_0(\underline{\mathbf{x}}^k)$ and continued with **II**.

II: $\underline{\mathbf{d}}^k = -\left(\mathbf{I} - {\mathbf{A}_k}^{\mathrm{T}}(\mathbf{A}_k{\mathbf{A}_k}^{\mathrm{T}})^{-1}\mathbf{A}_k\right)\nabla f(\underline{\mathbf{x}}^k)$ is sunstituted. If $\underline{\mathbf{d}}^k \neq \underline{\mathbf{0}}$, then continued with **III**. If $\underline{\mathbf{d}}^k = 0$ and $\underline{\mathbf{u}} = -(\mathbf{A}_k{\mathbf{A}_k}^{\mathrm{T}})^{-1}\mathbf{A}_k\nabla f(\underline{\mathbf{x}}^k) \geq 0$, then $\underline{\mathbf{x}}^k$ is a minimum point. The local Kuhn-Tucker conditions $-\nabla f(\underline{\mathbf{x}}^k) = \sum\limits_{i\in I_0(\underline{\mathbf{x}}^k)} u_i\underline{\mathbf{a}}_i = {\mathbf{A}_k}^{\mathrm{T}}\underline{\mathbf{u}}$ are obviously satisfied.

If $\underline{\mathbf{u}} \not\geq \underline{\mathbf{0}}$, then an i with $u_i < 0$ is chosen, the i-th row from $\mathbf{A}_k$ is deleted and **II** is repeated.

III: Calculation of α_k and $\underline{\mathbf{x}}^{k+1} = \underline{\mathbf{x}}^k + \alpha_k\underline{\mathbf{d}}^k$ and returning to **I** with $k = k + 1$.

3. Remarks on the Algorithm

If $-\nabla f(\underline{\mathbf{x}}^k)$ is not feasible, then this vector is mapped onto the sub-manifold of the smallest dimension which contains $\underline{\mathbf{x}}^k$. If $\underline{\mathbf{d}}^k = \underline{\mathbf{0}}$, then $-\nabla f(\underline{\mathbf{x}}^k)$ is perpendicular to this sub-manifold. If $\underline{\mathbf{u}} \geq \underline{\mathbf{0}}$ does not hold, then the dimension of the sub-manifold is increased by one by omitting one of the active constraints, so maybe $\underline{\mathbf{d}}^k \neq \underline{\mathbf{0}}$ can occur **(Fig. 18.10b)** (with projection onto a (lateral) face). Since $\mathbf{A}_k$ is often obtained from $\mathbf{A}_{k-1}$ by adding or erasing a row, the calculation of $(\mathbf{A}_k{\mathbf{A}_k}^{\mathrm{T}})^{-1}$ can be simplified by the use of $(\mathbf{A}_{k-1}{\mathbf{A}_{k-1}}^{\mathrm{T}})^{-1}$.

■ Solution of the problem of the previous example on p. 938.

Step 1: $\underline{\mathbf{x}}^1 = (3, 0)^{\mathrm{T}}$,

I: $\nabla f(\underline{\mathbf{x}}^1) = (-4, -32)^{\mathrm{T}}$, $-\nabla f(\underline{\mathbf{x}}^1)$ is feasible, $\underline{\mathbf{d}}^1 = (4, 32)^{\mathrm{T}}$.

III: The step size is determined as in the previous example: $\alpha_1 = \dfrac{1}{20}$, $\underline{\mathbf{x}}^2 = \left(\dfrac{16}{5}, \dfrac{8}{5}\right)^{\mathrm{T}}$.

Step 2:

I: $\nabla f(\underline{\mathbf{x}}^2) = \left(-\dfrac{18}{5}, -\dfrac{96}{5}\right)^{\mathrm{T}}$ (not feasible), $I_0(\underline{\mathbf{x}}^2) = \{4\}$, $\mathbf{A}_2 = (2\ 1)$.

II: $\mathbf{P}_2 = \dfrac{1}{5}\begin{pmatrix} 1 & -2 \\ -2 & 4 \end{pmatrix}$, $\underline{\mathbf{d}}^2 = \left(-\dfrac{8}{25}, \dfrac{16}{25}\right)^{\mathrm{T}} \neq \underline{\mathbf{0}}$.

III: $\alpha_2 = \frac{5}{8}$, $\underline{\mathbf{x}}^3 = (3,2)^{\mathrm{T}}$.

Step 3:

I: $\nabla f(\underline{\mathbf{x}}^3) = (-4,-16)^{\mathrm{T}}$ (not feasible), $I_0(\underline{\mathbf{x}}^3) = \{3,4\}$, $\mathbf{A}_3 = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$.

II: $\mathbf{P}_3 = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$, $\underline{\mathbf{d}}^3 = (0,0)^{\mathrm{T}}, \underline{\mathbf{u}} = \left(\frac{28}{3}, -\frac{8}{3}\right)^{\mathrm{T}}$ $u_2 < 0$: $\mathbf{A}_3 = (1\ 2)$.

II: $\mathbf{P}_3 = \frac{1}{5}\begin{pmatrix} 4 & -2 \\ -2 & 1 \end{pmatrix}$, $\underline{\mathbf{d}}^3 = \left(-\frac{16}{5}, \frac{8}{5}\right)^{\mathrm{T}}$.

III: $\alpha_3 = \frac{5}{16}$, $\underline{\mathbf{x}}^4 = \left(2, \frac{5}{2}\right)^{\mathrm{T}}$.

Step 4:

I: $\nabla f(\underline{\mathbf{x}}^4) = (-6,-12)^{\mathrm{T}}$ (not feasible), $I_0(\underline{\mathbf{x}}^4) = \{3\}$, $\mathbf{A}_4 = \mathbf{A}_3$.

II: $\mathbf{P}_4 = \mathbf{P}_3$, $\underline{\mathbf{d}}^4 = (0,0)^{\mathrm{T}}$, $u = 6 \geq 0$.

It follows that $\underline{\mathbf{x}}^4$ is a minimum point.

18.2.8 Penalty Function and Barrier Methods

The basic principle of these methods is that a constrained optimization problem is transformed into a sequence of optimization problems without constraints by modifying the objective function. The modified problem can be solved, e.g., by the methods given in Section 18.2.5. With an appropriate construction of the modified objective functions, every accumulation point of the sequence of the solution points of this modified problem is a solution of the original problem.

18.2.8.1 Penalty Function Method

The problem

$$f(\underline{\mathbf{x}}) = \min! \text{ subject to } g_i(\underline{\mathbf{x}}) \leq 0 \ \ (i = 1,2,\ldots,m) \tag{18.107}$$

is replaced by the sequence of unconstrained problems

$$H(\underline{\mathbf{x}}, p_k) = f(\underline{\mathbf{x}}) + p_k S(\underline{\mathbf{x}}) = \min! \text{ with } \underline{\mathbf{x}} \in \mathbb{R}^n, \ \ p_k > 0 \ \ (k = 1,2,\ldots). \tag{18.108}$$

Here, p_k is a positive parameter, and for $S(\underline{\mathbf{x}})$

$$S(\underline{\mathbf{x}}) = \begin{cases} = 0 & \underline{\mathbf{x}} \in M, \\ > 0 & \underline{\mathbf{x}} \notin M, \end{cases} \tag{18.109}$$

holds, i.e., leaving the feasible set M is punished with a "penalty" term $p_k S(\underline{\mathbf{x}})$. The problem (18.108) is solved with a sequence of penalty parameters p_k tending to ∞. Then

$$\lim_{k\to\infty} H(\underline{\mathbf{x}}, p_k) = f(\underline{\mathbf{x}}), \quad \underline{\mathbf{x}} \in M. \tag{18.110}$$

If $\underline{\mathbf{x}}^k$ is the solution of the k-th penalty problem, then:

$$H(\underline{\mathbf{x}}^k, p_k) \geq H(\underline{\mathbf{x}}^{k-1}, p_{k-1}), \quad f(\underline{\mathbf{x}}^k) \geq f(\underline{\mathbf{x}}^{k-1}), \tag{18.111}$$

and every accumulation point $\underline{\mathbf{x}}^*$ of the sequence $\{\underline{\mathbf{x}}^k\}$ is a solution of (18.107). If $\underline{\mathbf{x}}^k \in M$, then $\underline{\mathbf{x}}^k$ is a solution of the original problem.

For instance, the following functions are appropriate realizations of $S(\underline{\mathbf{x}})$:

$$S(\underline{\mathbf{x}}) = \max{}^r\{0, g_1(\underline{\mathbf{x}}), \ldots, g_m(\underline{\mathbf{x}})\} \ \ (r = 1,2,\ldots) \text{ or} \tag{18.112a}$$

$$S(\underline{\mathbf{x}}) = \sum_{i=1}^{m} \max{}^r\{0, g_i(\underline{\mathbf{x}})\} \ \ (r = 1,2,\ldots). \tag{18.112b}$$

If functions $f(\underline{\mathbf{x}})$ and $g_i(\underline{\mathbf{x}})$ are differentiable, then in the case $r > 1$, the penalty function $H(\underline{\mathbf{x}}, p_k)$ is also differentiable on the boundary of M, so analytic solutions can be used to solve the auxiliary problem (18.108).

Fig. 18.11 shows a representation of the penalty function method.

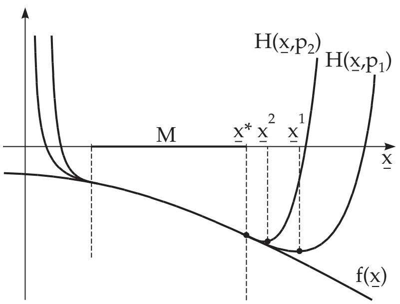


Figure 18.11

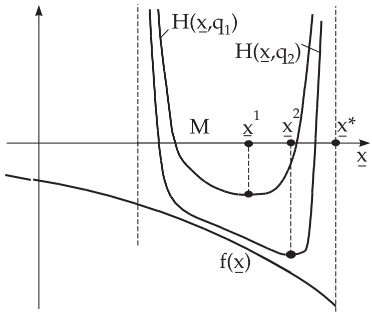


Figure 18.12

■ $f(\underline{\mathbf{x}}) = x_1^2 + x_2^2 =$ min! for $x_1 + x_2 \geq 1,\ H(\underline{\mathbf{x}}, p_k) = x_1^2 + x_2^2 + p_k \max^2\{0, 1 - x_1 - x_2\}$.
The necessary optimality condition is:
$\nabla H(\underline{\mathbf{x}}, p_k) = \begin{pmatrix} 2x_1 - 2p_k \max\{0, 1 - x_1 - x_2\} \\ 2x_2 - 2p_k \max\{0, 1 - x_1 - x_2\} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$.
The gradient of H is evaluated here with respect to $\underline{\mathbf{x}}$. By subtracting the equations we have $x_1 = x_2$. The equation $2x_1 - 2p_k \max\{0, 1 - 2x_1\} = 0$ has a unique solution $x_1^k = x_2^k = \dfrac{p_k}{1 + 2p_k}$. We get the solution $x_1^* = x_2^* = \lim\limits_{k\to\infty} \dfrac{p_k}{1 + 2p_k} = \dfrac{1}{2}$ by letting $k \to \infty$.

18.2.8.2 Barrier Method

A sequence of modified problems is considered in the form

$$H(\underline{\mathbf{x}}, q_k) = f(\underline{\mathbf{x}}) + q_k B(\underline{\mathbf{x}}) = \text{min!}\,, \quad q_k > 0\,. \tag{18.113}$$

The term $q_k B(\underline{\mathbf{x}})$ prevents the solution leaving the feasible set M, since the objective function increases unboundedly on approaching the boundary of M. The *regularity condition*

$$M^0 = \{\underline{\mathbf{x}} \in M \ :\ g_i(\underline{\mathbf{x}}) < 0 \ \ (i = 1, 2, \ldots, m)\} \neq \emptyset \quad \text{and} \quad \overline{M^0} = M \tag{18.114}$$

must be satisfied, i.e., the interior of M must be non-empty and it is possible to get to any boundary point by approaching it from the interior, i.e., the closure of M^0 is M.
The function $B(\underline{\mathbf{x}})$ is defined to be continuous on M^0. It increases to ∞ at the boundary of M. The modified problem (18.113) is solved by a sequence of barrier parameters q_k tending to zero. For the solution $\underline{\mathbf{x}}^k$ of the k-th problem (18.113)

$$f(\underline{\mathbf{x}}^k) \leq f(\underline{\mathbf{x}}^{k-1}), \tag{18.115}$$

holds and every accumulation point $\underline{\mathbf{x}}^*$ of the sequence $\{\underline{\mathbf{x}}^k\}$ is a solution of (18.107).
Fig. 18.12 shows a representation of the barrier method.
The functions, e.g.,

$$B(\underline{\mathbf{x}}) = -\sum_{i=1}^{m} - \ln(-g_i(\underline{\mathbf{x}})), \quad \underline{\mathbf{x}} \in M^0 \quad \text{or} \tag{18.116a}$$

$$B(\underline{\mathbf{x}}) = \sum_{i=1}^{m} \frac{1}{[-g_i(\underline{\mathbf{x}})]^r} \quad (r = 1, 2, \ldots), \quad \underline{\mathbf{x}} \in M^0 \tag{18.116b}$$

are appropriate realizations of $B(\underline{\mathbf{x}})$.

■ $f(\underline{\mathbf{x}}) = x_1^2 + x_2^2 = \min!$ subject to $x_1 + x_2 \geq 1$, $H(\underline{\mathbf{x}}, q_k) = x_1^2 + x_2^2 + q_k(-\ln(x_1 + x_2 - 1))$, $x_1 + x_2 > 1$, $\nabla H(\underline{\mathbf{x}}, q_k) = \begin{pmatrix} 2x_1 - q_k \dfrac{1}{x_1 - x_2 - 1} \\ 2x_2 - q_k \dfrac{1}{x_1 + x_2 - 1} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, $x_1 + x_2 > 1$. The gradient of H is given with respect to $\underline{\mathbf{x}}$.

Subtracting the equations results in $x_1 = x_2$, $2x_1 - q_k \dfrac{1}{2x_1 - 1} = 0$, $x_1 > \dfrac{1}{2}$, $\Longrightarrow$ $x_1^2 - \dfrac{x_1}{2} - \dfrac{q_k}{4} = 0$, $x_1 > \dfrac{1}{2}$, $x_1^k = x_2^k = \dfrac{1}{4} + \sqrt{\dfrac{1}{16} + \dfrac{1}{4}q_k}$, $k \to \infty$, $q_k \to 0$: $x_1^* = x_2^* = \dfrac{1}{2}$.

The solutions of problems (18.108) and (18.113) at the k-th step do not depend on the solutions of the previous steps. The application of higher penalty or smaller barrier parameters often leads to convergence problems with numerical solutions of (18.108) and (18.113), e.g., in the method of (18.2.4), in particular, if no good initial approximation is available. Using the result of the k-th problem as the initial solution for the $(k+1)$-th problem the convergence behavior can be improved.

18.2.9 Cutting Plane Methods

1. Formulation of the Problem and Principle of Solution

Let consider the problem

$$f(\underline{\mathbf{x}}) = \underline{\mathbf{c}}^{\mathrm{T}}\underline{\mathbf{x}} = \min!, \quad \underline{\mathbf{c}} \in \mathbb{R}^n \tag{18.117}$$

over the bounded region $M \subset \mathbb{R}^n$ given by convex functions $g_i(\underline{\mathbf{x}})$ $(i = 1, 2, \dots, m)$ in the form $g_i(\underline{\mathbf{x}}) \leq 0$. A problem with a non-linear but convex objective function $f(\underline{\mathbf{x}})$ is transformed into this form, if

$$f(\underline{\mathbf{x}}) - x_{n+1} \leq 0, \quad x_{n+1} \in \mathbb{R} \tag{18.118}$$

is considered as a further constraint and

$$\overline{f}(\overline{\underline{\mathbf{x}}}) = x_{n+1} = \min! \quad \text{for all} \quad \overline{\underline{\mathbf{x}}} = (\underline{\mathbf{x}}, x_{n+1}) \in \mathbb{R}^{n+1} \tag{18.119}$$

is solved with $\overline{g}_i(\overline{\underline{\mathbf{x}}}) = g_i(\underline{\mathbf{x}}) \leq 0$.

The basic idea of the method is the iterative linear approximation of M by a convex polyhedron in the neighborhood of the minimum point $\underline{\mathbf{x}}^*$, and therefore the original program is reduced to a sequence of linear programming problems.

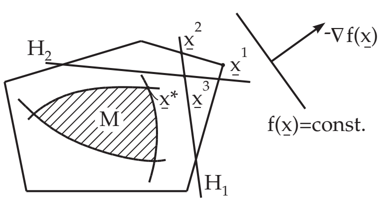


Figure 18.13

First, a polyhedron is determined:

$$P_1 = \{\underline{\mathbf{x}} \in \mathbb{R}^n \; : \; \underline{\mathbf{a}}_i{}^{\mathrm{T}}\underline{\mathbf{x}} \leq b_i, \;\; i = 1, \dots, s\}. \tag{18.120}$$

From the linear program

$$f(\underline{\mathbf{x}}) = \min! \quad \text{with} \quad \underline{\mathbf{x}} \in P_1 \tag{18.121}$$

an optimal extreme point $\underline{\mathbf{x}}^1$ of P_1 is determined with respect to $f(\underline{\mathbf{x}})$. If $\underline{\mathbf{x}}^1 \in M$ holds, then the optimal solution of the original problem is found. Otherwise, a hyperplane is determined:

$H_1 = \{\underline{\mathbf{x}} : \; \underline{\mathbf{a}}_{s+1}{}^{\mathrm{T}}\underline{\mathbf{x}} = b_{s+1}, \; \underline{\mathbf{a}}_{s+1}{}^{\mathrm{T}}x^1 > b_{s+1}\}$, which separates the point $\underline{\mathbf{x}}^1$ from M, so the new polyhedron contains

$$P_2 = \{\underline{\mathbf{x}} \in P_1 : \; \underline{\mathbf{a}}_{s+1}{}^{\mathrm{T}}\underline{\mathbf{x}} \leq b_{s+1}\}. \tag{18.122}$$

Fig. 18.13 shows a schematic representation of the cutting plane method.

2. Kelley Method

The different methods differ from each other in the choice of the separating planes H_k. In the case of the Kelley method H_k is chosen in the following way: A j_k is chosen so that

$$g_{j_k}(\underline{\mathbf{x}}^k) = \max\{g_i(\underline{\mathbf{x}}^k) \;\; (i = 1, \dots, m)\}. \tag{18.123}$$

At the point $\underline{\mathbf{x}} = \underline{\mathbf{x}}^k$, the function $g_{jk}(\underline{\mathbf{x}})$ has the tangent plane

$$T(\underline{\mathbf{x}}) = g_{j_k}(\underline{\mathbf{x}}^k) + (\underline{\mathbf{x}} - \underline{\mathbf{x}}^k)^{\mathrm{T}} \nabla g_{j_k}(\underline{\mathbf{x}}^k). \tag{18.124}$$

The hyperplane $H_k = \{\underline{\mathbf{x}} \in \mathbb{R}^n \ : \ T(\underline{\mathbf{x}}) = 0\}$ separates the point $\underline{\mathbf{x}}^k$ from all points $\underline{\mathbf{x}}$ with $g_{j_k}(\underline{\mathbf{x}}) \le 0$. So, for the $(k+1)$-th linear program, $T(\underline{\mathbf{x}}) \le 0$ is added as a further constraint. Every accumulation point $\underline{\mathbf{x}}^*$ of the sequence $\{\underline{\mathbf{x}}^k\}$ is a minimum point of the original problem.
In practical applications this method shows a low speed of convergence. Furthermore, the number of constraints is always increasing.

18.3 Discrete Dynamic Programming

18.3.1 Discrete Dynamic Decision Models

A wide class of optimization problems can be solved by the methods of dynamic programming. The problem is considered as a *process* proceeding naturally or formally in time, and it is controlled by time-dependent decisions. If the process can be decomposed into a finite or countably infinite number of steps, then it is called *discrete dynamic programming*, otherwise it is a *continuous dynamic programming*. In this section, only n-stage discrete processes are discussed.

18.3.1.1 n-Stage Decision Processes

An n-stage process P starts at stage 0 with an initial state $\underline{\mathbf{x}}_a = \underline{\mathbf{x}}_0$ and proceeds through the intermediate states $\underline{\mathbf{x}}_1, \underline{\mathbf{x}}_2, \ldots, \underline{\mathbf{x}}_{n-1}$ into a final state $\underline{\mathbf{x}}_n = \underline{\mathbf{x}}_e \in X_e \subseteq \mathbb{R}^m$. The *state vectors* $\underline{\mathbf{x}}_j$ are in the state space $X_j \subseteq \mathbb{R}^m$. To drive a state $\underline{\mathbf{x}}_{j-1}$ into the state $\underline{\mathbf{x}}_j$, a *decision* $\underline{\mathbf{u}}_j$ is required. All possible *decision vectors* $\underline{\mathbf{u}}_j$ in the state $\underline{\mathbf{x}}_{j-1}$ form the *decision space* $U_j(\underline{\mathbf{x}}_{j-1}) \subseteq \mathbb{R}^s$. From $\underline{\mathbf{x}}_{j-1}$ the consecutive state $\underline{\mathbf{x}}_j$ can be obtained by the transformation **(Fig. 18.14)**

$$\underline{\mathbf{x}}_j = g_j(\underline{\mathbf{x}}_{j-1}, \underline{\mathbf{u}}_j), \quad j = 1(1)n. \tag{18.125}$$

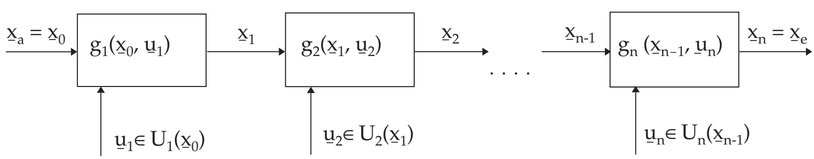


Figure 18.14

18.3.1.2 Dynamic Programming Problem

Our goal is to determine a *policy* $(\underline{\mathbf{u}}_1, \ldots, \underline{\mathbf{u}}_n)$ which drives the process from the initial state $\underline{\mathbf{x}}_a$ into the state $\underline{\mathbf{x}}_e$ considering all constraints so that it minimizes an objective function or *cost function* $f(f_1(\underline{\mathbf{x}}_0, \underline{\mathbf{u}}_1), \ldots, f_n(\underline{\mathbf{x}}_{n-1}, \underline{\mathbf{u}}_n))$. The functions $f_j(\underline{\mathbf{x}}_{j-1}, \underline{\mathbf{u}}_j)$ are called *stage costs*. The standard form of the dynamic programming problem is

OF: $$f(f_1(\underline{\mathbf{x}}_0, \underline{\mathbf{u}}_1), \ldots, f_n(\underline{\mathbf{x}}_{n-1}, \underline{\mathbf{u}}_n)) \longrightarrow \min! \tag{18.126a}$$

CT: $$\left.\begin{array}{ll} \underline{\mathbf{x}}_j = g_j(\underline{\mathbf{x}}_{j-1}, \underline{\mathbf{u}}_j), & j = 1(1)n, \\ \underline{\mathbf{x}}_0 = \underline{\mathbf{x}}_a,\ \underline{\mathbf{x}}_n = \underline{\mathbf{x}}_e \in X_e,\ \underline{\mathbf{x}}_j \in X_j \subseteq \mathbb{R}^m, & j = 1(1)n, \\ \underline{\mathbf{u}}_j \in U_j(\underline{\mathbf{x}}_{j-1}) \subseteq \mathbb{R}^m, & j = 1(1)n. \end{array}\right\} \tag{18.126b}$$

The first type of constraints $\underline{\mathbf{x}}_j$ are called *dynamic* and the others $\underline{\mathbf{x}}_0, \underline{\mathbf{u}}_j$ are called *static*. Similarly to (18.126a), a maximum problem can also be considered. A policy $(\underline{\mathbf{u}}_1, \ldots, \underline{\mathbf{u}}_n)$ satisfying all constraints is called *feasible*. The methods of dynamic programming can be applied if the objective function satisfies certain additional requirements (see 18.3.3, p. 944).

18.3.2 Examples of Discrete Decision Models

18.3.2.1 Purchasing Problem

In the j-th period of a time interval which can be divided into n periods, a workshop needs v_j units of a certain primary material. The available amount of this material at the beginning of period j is denoted by x_{j-1}, in particular, $x_0 = x_a$ is given. The amounts u_j are to be determined for a unit price c_j, which should be purchased by the workshop at the end of each period. The given storage capacity K must not be exceeded, i.e., $x_{j-1} + u_j \le K$. A purchase policy $(u_1, \dots, u_n)$ should be determined, which minimizes the total cost. This problem leads to the following dynamic programming problem:

OF: $$f(u_1, \dots, u_n) = \sum_{j=1}^{n} f_j(u_j) = \sum_{j=1}^{n} c_j u_j \longrightarrow \min! \tag{18.127a}$$

CT: $$\left.\begin{array}{ll} x_j = x_{j-1} + u_j - v_j\,, & j = 1(1)n, \\ x_0 = x_a\,, \quad 0 \le x_j \le K\,, & j = 1(1)n, \\ U_j(x_{j-1}) = \{u_j : \max\{0, v_j - x_{j-1}\} \le u_j \le K - x_{j-1}\}\,, & j = 1(1)n. \end{array}\right\} \tag{18.127b}$$

In (18.127b) it is ensured that demands are satisfied and the storage capacity is not exceeded. If there is also a storage cost l per unit per period, then intermediate cost in the j-th period is $(x_{j-1} + u_j - v_j/2)l$, and the modified cost function is

$$f(x_0, u_1, \dots, x_{n-1}, u_n) = \sum_{j=1}^{n} (c_j u_j + (x_{j-1} + u_j - v_j/2) \cdot l). \tag{18.128}$$

18.3.2.2 Knapsack Problem

One has to select some of the items $A_1, \dots, A_n$ with weights $w_1, \dots, w_n$ and with values $c_1, \dots, c_n$ so that the total weight does not exceed a given bound W, and the total value is maximal. This problem does not depend on time. It will be reformulated in the following way: At every stage a decision u_j about the selection of item A_j is made. Here, $u_j = 1$ holds if A_j is selected, otherwise $u_j = 0$. The capacity still available at the beginning of a stage is denoted by x_{j-1}, so the following dynamic problem arises:

OF: $$f(u_1, \dots, u_n) = \sum_{j=1}^{n} c_j u_j \longrightarrow \max! \tag{18.129a}$$

CT: $$\left.\begin{array}{ll} x_j = x_{j-1} - w_j u_j, & j = 1(1)n, \\ x_0 = W\,, \quad 0 \le x_j \le W, & j = 1(1)n, \\ \left.\begin{array}{ll} u_j \in \{0,1\}, & \text{falls } x_{j-1} \ge w_j, \\ u_j = 0\,, & \text{falls } x_{j-1} < w_j, \end{array}\right\} & j = 1(1)n. \end{array}\right\} \tag{18.129b}$$

18.3.3 Bellman Functional Equations

18.3.3.1 Properties of the Cost Function

In order to state the Bellman functional equations, the cost function must satisfy two requirements:

1. Separability

The function $f(f_1(\underline{\mathbf{x}}_0, \underline{\mathbf{u}}_1), \dots, f_n(\underline{\mathbf{x}}_{n-1}, \underline{\mathbf{u}}_n))$ is called *separable*, if it can be given by binary functions $H_1, \dots, H_{n-1}$ and by functions $F_1, \dots, F_n$ in the following way:

$$\begin{array}{l} f(f_1(\underline{\mathbf{x}}_0, \underline{\mathbf{u}}_1), \dots, f_n(\underline{\mathbf{x}}_{n-1}, \underline{\mathbf{u}}_n)) = F_1(f_1(\underline{\mathbf{x}}_0, \underline{\mathbf{u}}_1), \dots, f_n(\underline{\mathbf{x}}_{n-1}, \underline{\mathbf{u}}_n)), \\ F_1(f_1(\underline{\mathbf{x}}_0, \underline{\mathbf{u}}_1), \dots, f_n(\underline{\mathbf{x}}_{n-1}, \underline{\mathbf{u}}_n)) = H_1\left(f_1(\underline{\mathbf{x}}_0, \underline{\mathbf{u}}_1), F_2(f_2(\underline{\mathbf{x}}_1, \underline{\mathbf{u}}_2), \dots, f_n(\underline{\mathbf{x}}_{n-1}, \underline{\mathbf{u}}_n))\right), \\ \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\ F_{n-1}(f_{n-1}(\underline{\mathbf{x}}_{n-2}, \underline{\mathbf{u}}_{n-1}), f_n(\underline{\mathbf{x}}_{n-1}, \underline{\mathbf{u}}_n)) = H_{n-1}\left(f_{n-1}(\underline{\mathbf{x}}_{n-2}, \underline{\mathbf{u}}_{n-1}), F_n(f_n(\underline{\mathbf{x}}_{n-1}, \underline{\mathbf{u}}_n))\right), \\ F_n(f_n(\underline{\mathbf{x}}_{n-1}, \underline{\mathbf{u}}_n)) = f_n(\underline{\mathbf{x}}_{n-1}, \underline{\mathbf{u}}_n). \end{array} \tag{18.130}$$

2. Minimum Interchangeability

A function $H(\tilde{f}(\underline{\mathbf{a}}), \tilde{F}(\underline{\mathbf{b}}))$ is called *minimum interchangeable*, if:

$$\min_{(\underline{\mathbf{a}},\underline{\mathbf{b}})\in A\times B} H\left(\tilde{f}(\underline{\mathbf{a}}), \tilde{F}(\underline{\mathbf{b}})\right) = \min_{\underline{\mathbf{a}}\in A} H\left(\tilde{f}(\underline{\mathbf{a}}), \min_{\underline{\mathbf{b}}\in B} \tilde{F}(\underline{\mathbf{b}})\right). \tag{18.131}$$

This property is satisfied, for example, if H is monotone increasing with respect to its second argument for every $\underline{\mathbf{a}} \in A$, i.e., if for every $\underline{\mathbf{a}} \in A$,

$$H\left(\tilde{f}(\underline{\mathbf{a}}), \tilde{F}(\underline{\mathbf{b}}_1)\right) \le H\left(\tilde{f}(\underline{\mathbf{a}}), \tilde{F}(\underline{\mathbf{b}}_2)\right) \quad \text{for} \quad \tilde{F}(\underline{\mathbf{b}}_1) \le \tilde{F}(\underline{\mathbf{b}}_2). \tag{18.132}$$

Now, for the cost function of the dynamic programming problem, the separability of f and the minimum interchangeability of all functions H_j, $j = 1(1)n-1$ are required. The following often occurring type of cost function satisfies both requirements:

$$f^{sum} = \sum_{j=1}^{n} f_j(\underline{\mathbf{x}}_{j-1}, \underline{\mathbf{u}}_j) \quad \text{or} \quad f^{max} = \max_{j=1(1)n} f_j(\underline{\mathbf{x}}_{j-1}, \underline{\mathbf{u}}_j). \tag{18.133}$$

The functions H_j are

$$H_j^{sum} = f_j(\underline{\mathbf{x}}_{j-1}, \underline{\mathbf{u}}_j) + \sum_{k=j+1}^{n} f_k(\underline{\mathbf{x}}_{k-1}, \underline{\mathbf{u}}_k) \quad \text{and} \tag{18.134}$$

$$H_j^{max} = \max\left\{ f_j(\underline{\mathbf{x}}_{j-1}, \underline{\mathbf{u}}_j), \max_{k=j+1(1)n} f_k(\underline{\mathbf{x}}_{k-1}, \underline{\mathbf{u}}_k)\right\}. \tag{18.135}$$

18.3.3.2 Formulation of the Functional Equations

The following functions are defined:

$$\phi_j(\underline{\mathbf{x}}_{j-1}) = \min_{\substack{\underline{\mathbf{u}}_k\in U_k(\underline{\mathbf{x}}_{k-1})\\ k=j(1)n}} F_j(f_j(\underline{\mathbf{x}}_{j-1}, \underline{\mathbf{u}}_j), \dots, f_n(\underline{\mathbf{x}}_{n-1}, \underline{\mathbf{u}}_n)), \quad j = 1(1)n, \tag{18.136}$$

$$\phi_{n+1}(\underline{\mathbf{x}}_n) = 0. \tag{18.137}$$

If there is no policy $(\underline{\mathbf{u}}_1, \dots, \underline{\mathbf{u}}_n)$ driving the state $\underline{\mathbf{x}}_{j-1}$ into a final state $\underline{\mathbf{x}}_e \in X_e$, then we substitute $\phi_j(\underline{\mathbf{x}}_{j-1}) = \infty$. Using the separability and minimum interchangeability conditions and the dynamic constraints for $j = 1(1)n$ we get:

$$\begin{aligned}
\phi_j(\underline{\mathbf{x}}_{j-1}) &= \min_{\underline{\mathbf{u}}_j\in U_j(\underline{\mathbf{x}}_{j-1})} H_j(f_j(\underline{\mathbf{x}}_{j-1}, \underline{\mathbf{u}}_j), \min_{\substack{\underline{\mathbf{u}}_k\in U_k(\underline{\mathbf{x}}_{k-1})\\ k=j+1(1)n}} F_{j+1}(f_{j+1}(\underline{\mathbf{x}}_j, \underline{\mathbf{u}}_{j+1}), \dots, f_n(\underline{\mathbf{x}}_{n-1}, \underline{\mathbf{u}}_n))),\\
&= \min_{\underline{\mathbf{u}}_j\in U_j(\underline{\mathbf{x}}_{j-1})} H_j\left(f_j(\underline{\mathbf{x}}_{j-1}, \underline{\mathbf{u}}_j), \phi_{j+1}(\underline{\mathbf{x}}_j)\right)
\end{aligned}$$

$$\phi_j(\underline{\mathbf{x}}_{j-1}) = \min_{\underline{\mathbf{u}}_j\in U_j(\underline{\mathbf{x}}_{j-1})} H_j\left(f_j(\underline{\mathbf{x}}_{j-1}, \underline{\mathbf{u}}_j), \phi_{j+1}(g_j(\underline{\mathbf{x}}_{j-1}, \underline{\mathbf{u}}_j))\right). \tag{18.138}$$

Equations (18.138) and (18.137) are called the *Bellman functional equations.* $\phi_1(\underline{\mathbf{x}}_0)$ is the optimal value of the cost function f.

18.3.4 Bellman Optimality Principle

The evaluation of the functional equation

$$\phi_j(\underline{\mathbf{x}}_{j-1}) = \min_{\underline{\mathbf{u}}_j\in U_j(\underline{\mathbf{x}}_{j-1})} H_j\left(f_j(\underline{\mathbf{x}}_{j-1}, \underline{\mathbf{u}}_j), \phi_{j+1}(\underline{\mathbf{x}}_j)\right) \tag{18.139}$$

corresponds to the determination of an optimal policy $(\underline{\mathbf{u}}_j^*, \ldots, \underline{\mathbf{u}}_n^*)$ for which the subprocess P_j starting at state $\underline{\mathbf{x}}_{j-1}$ and consisting of the last $n-j+1$ stages of the total process P minimizes the cost function, i.e.,

$$F_j(f_j(\underline{\mathbf{x}}_{j-1}, \underline{\mathbf{u}}_j), \ldots, f_n(\underline{\mathbf{x}}_{n-1}, \underline{\mathbf{u}}_n)) \longrightarrow \text{min!}. \tag{18.140}$$

The optimal policy of the process P_j with the initial state $\underline{\mathbf{x}}_{j-1}$ is independent of the decisions $\underline{\mathbf{u}}_1, \ldots, \underline{\mathbf{u}}_{j-1}$ of the first $j-1$ stages of P which have driven P to the state $\underline{\mathbf{x}}_{j-1}$. To determine $\phi_j(\underline{\mathbf{x}}_{j-1})$ the value $\phi_{j+1}(\underline{\mathbf{x}}_j)$ is needed to know. Now, if $(\underline{\mathbf{u}}_j^*, \ldots, \underline{\mathbf{u}}_n^*)$ is an optimal policy for P_j, then, obviously, $(\underline{\mathbf{u}}_{j+1}^*, \ldots, \underline{\mathbf{u}}_n^*)$ is an optimal policy for the subprocess P_{j+1} starting at $\underline{\mathbf{x}}_j = g_j(\underline{\mathbf{x}}_{j-1}, \underline{\mathbf{u}}_j^*)$. This statement is generalized in the *Bellman optimality principle*.

Bellman Principle: If $(\underline{\mathbf{u}}_1^*, \ldots, \underline{\mathbf{u}}_n^*)$ is an optimal policy of the process P and $(\underline{\mathbf{x}}_0^*, \ldots, \underline{\mathbf{x}}_n^*)$ is the corresponding sequence of states, then for every subprocess $P_j, j = 1(1)n$, with initial state $\underline{\mathbf{x}}_{j-1}^*$ the policy $(\underline{\mathbf{u}}_j^*, \ldots, \underline{\mathbf{u}}_n^*)$ is also optimal.

18.3.5 Bellman Functional Equation Method

18.3.5.1 Determination of Minimal Costs

With the functional equations (18.137), (18.138) and starting with $\phi_{n+1}(\underline{\mathbf{x}}_n) = 0$ every value $\phi_j(\underline{\mathbf{x}}_{j-1})$ with $\underline{\mathbf{x}}_{j-1} \in X_{j-1}$ is determined in decreasing order of j. It requires the solution of an optimum problem over the decision space $U_j(\underline{\mathbf{x}}_{j-1})$ for every $\underline{\mathbf{x}}_{j-1} \in X_{j-1}$. For every $\underline{\mathbf{x}}_{j-1}$ there is a minimum point $\underline{\mathbf{u}}_j \in U_j$ as an optimal decision for the first stage of a subprocess P_j starting at $\underline{\mathbf{x}}_{j-1}$. If the sets X_j are not finite or they are too large, then the values ϕ_j can be calculated for a set of selected nodes $\underline{\mathbf{x}}_{j-1} \in X_{j-1}$. The intermediate values can be calculated by a certain interpolation method. $\phi_1(\underline{\mathbf{x}}_0)$ is the optimal value of the cost function of process P. The optimal policy $(\underline{\mathbf{u}}_1^*, \ldots, \underline{\mathbf{u}}_n^*)$ and the corresponding states $(\underline{\mathbf{x}}_0^*, \ldots, \underline{\mathbf{x}}_n^*)$ can be determined by one of the following two methods.

18.3.5.2 Determination of the Optimal Policy

1. Variant 1: During the evaluation of the functional equations, the computed $\underline{\mathbf{u}}_j$ is also saved for every $\underline{\mathbf{x}}_{j-1} \in X_{j-1}$. After the calculation of $\phi_1(\underline{\mathbf{x}}_0)$, an optimal policy is got if $\underline{\mathbf{x}}_1^* = g_1(\underline{\mathbf{x}}_0^*, \underline{\mathbf{u}}_1^*)$ is determined from $\underline{\mathbf{x}}_0 = \underline{\mathbf{x}}_0^*$ and the saved $\underline{\mathbf{u}}_1 = \underline{\mathbf{u}}_1^*$. From $\underline{\mathbf{x}}_1^*$ and the saved decision $\underline{\mathbf{u}}_2^*$ follows $\underline{\mathbf{x}}_2^*$, etc.

2. Variant 2: For every $\underline{\mathbf{x}}_{j-1} \in X_{j-1}$ only $\phi_j(\underline{\mathbf{x}}_{j-1})$ is saved. After every $\phi_j(\underline{\mathbf{x}}_{j-1})$ is known, a forward calculation is made. Starting with $j = 1$ and $\underline{\mathbf{x}}_0 = \underline{\mathbf{x}}_0^*$ one determines $\underline{\mathbf{u}}_j^*$ in increasing order of j by the evaluation of the functional equation

$$\phi_j(\underline{\mathbf{x}}_{j-1}^*) = \min_{\underline{\mathbf{u}}_j \in U_j(\underline{\mathbf{x}}_{j-1}^*)} H_j\left(f_j(\underline{\mathbf{x}}_{j-1}^*, \underline{\mathbf{u}}_j), \phi_{j+1}(g_j(\underline{\mathbf{x}}_{j-1}^*, \underline{\mathbf{u}}_j))\right). \tag{18.141}$$

$\underline{\mathbf{x}}_j^* = g_j(\underline{\mathbf{x}}_{j-1}^*, \underline{\mathbf{u}}_j^*)$ is obtained. During the forward calculation, an optimization problem must be solved again at every stage.

3. Comparison of the two Variants: The computation costs of variant **1** are less than variant **2** requires because of the forward calculations. However decision $\underline{\mathbf{u}}_j$ is saved for every state $\underline{\mathbf{x}}_{j-1}$, which may require very large memory in the case of a higher dimensional decision space $U_j(\underline{\mathbf{x}}_{j-1})$, while in the case of variant **2**, only the values $\phi_j(\underline{\mathbf{x}}_{j-1})$ must be saved. Therefore, sometimes variant **2** is used on computers.

18.3.6 Examples for Applications of the Functional Equation Method

18.3.6.1 Optimal Purchasing Policy

1. Formulation of the Problem

The problem from 18.3.2.1, p. 944, to determine an optimal purchasing policy

$$\textbf{OF} \quad f(u_1,\ldots,u_n) = \sum_{j=1}^{n} c_j u_j \longrightarrow \min! \tag{18.142a}$$

$$\begin{aligned} \textbf{CT} \quad & x_j = x_{j-1} + u_j - v_j, \quad j = 1(1)n, \\ & x_0 = x_a,\ 0 \le x_j \le K, \quad j = 1(1)n, \\ & U_j(x_{j-1}) = \{u_j\ :\ \max\{0, v_j - x_{j-1}\} \le u_j \le K - x_{j-1}\}, \quad j = 1(1)n \end{aligned} \tag{18.142b}$$

leads to the functional equations

$$\phi_{n+1}(x_n) = 0, \tag{18.143}$$

$$\phi_j(x_{j-1}) = \min_{u_j \in U_j(x_{j-1})} \left(c_j u_j + \phi_{j+1}(x_{j-1} + u_j - v_j)\right), \quad j = 1(1)n. \tag{18.144}$$

2. Numerical Example

$n = 6, \quad K = 10, \quad x_a = 2.$ $\quad c_1 = 4, \quad c_2 = 3, \quad c_3 = 5, \quad c_4 = 3, \quad c_5 = 4, \quad c_6 = 2,$
$v_1 = 6, \quad v_2 = 7, \quad v_3 = 4, \quad v_4 = 2, \quad v_5 = 4, \quad v_6 = 3.$

Backward Calculation: The function values $\phi_j(x_{j-1})$ will be determined for the states $x_{j-1} = 0, 1, \ldots, 10$. Now, it is enough to make the minimum search only for integer values of u_j.

$$j = 6: \quad \phi_6(x_5) = \min_{u_6 \in U_6(x_5)} c_6 u_6 = c_6 \max\{0, v_6 - x_5\} = 2\max\{0, 3 - x_5\}.$$

According to variant **2** of the Bellman functional equation method, only the values of $\phi_6(x_5)$ are written in the last row. For example, $\phi_4(0)$ is determined.

$$\begin{aligned} \phi_4(0) &= \min_{2 \le u_4 \le 10} (3u_4 + \phi_5(u_4 - 2)) \\ &= \min(28, 27, 26, 25, 24, 25, 26, 27, 30) = 24. \end{aligned}$$

	x_j=0	1	2	3	4	5	6	7	8	9	10
j=1			75								
2	59	56	53	50	47	44	41	38	35	32	29
3	44	39	34	29	24	21	18	15	12	9	6
4	24	21	18	15	12	9	6	4	2	0	0
5	22	18	14	10	6	4	2	0	0	0	0
6	6	4	2	0	0	0	0	0	0	0	0

Forward Calculation:

$$\phi_1(2) = 75 = \min_{4 \le u_1 \le 8} (4u_1 + \phi_2(u_1 - 4)).$$

One gets $u_1^* = 4$ as the minimum point, therefore $x_1^* = x_0^* + u_1^* - v_1 = 0$. This method is repeated for $\phi_2(0)$ and for all later stages. The optimal policy is:

$$(u_1^*, u_2^*, u_3^*, u_4^*, u_5^*, u_6^*) = (4, 10, 1, 6, 0, 3).$$

18.3.6.2 Knapsack Problem

1. Formulation of the Problem

Consider the problem given in 18.3.2.2, p. 944

$$\textbf{OF}: f(u_1,\ldots,u_n) = \sum_{j=1}^{n} c_j u_j \longrightarrow \max! \tag{18.145a}$$

$$\textbf{CT:} \quad \left.\begin{aligned} & x_j = x_{j-1} - w_j u_j, && j = 1(1)n, \\ & x_0 = W, \quad 0 \le x_j \le W, && j = 1(1)n, \\ & \left.\begin{aligned} & u_j \in \{0, 1\}, \text{ if } x_{j-1} \ge w_j, \\ & u_j = 0, \qquad \text{if } x_{j-1} < w_j, \end{aligned}\right\} && j = 1(1)n. \end{aligned}\right\} \tag{18.145b}$$

Since this is a maximum problem, the Bellman functional equations are now

$$\phi_{n+1}(x_n) = 0,$$

$$\phi_j(x_{j-1}) = \max_{u_j \in U_j(x_{j-1})} (c_j u_j + \phi_{j+1}(x_{j-1} - w_j u_j)), \quad j = 1(1)n.$$

The decisions can be only 0 and 1, so it is practical to apply variant **1** of the functional equation method. For $j = n, n-1, \ldots, 1$:

$$\phi_j(x_{j-1}) = \begin{cases} c_j + \phi_{j+1}(x_{j-1} - w_j) & \text{if } x_{j-1} \geq w_j \text{ and } c_j + \phi_{j+1}(x_{j-1} - w_j) > \phi_{j+1}(x_{j-1}), \\ \phi_{j+1}(x_{j-1}) & \text{otherwise,} \end{cases}$$

$$u_j(x_{j-1}) = \begin{cases} 1 \text{ if } x_{j-1} \geq w_j \text{ and } c_j + \phi_{j+1}(x_{j-1} - w_j) > \phi_{j+1}(x_{j-1}), \\ 0 \text{ otherwise.} \end{cases}$$

2. Numerical Example

$$W = 10, \quad n = 6. \quad \begin{matrix} c_1 = 1, & c_2 = 2, & c_3 = 3, & c_4 = 1, & c_5 = 5, & c_6 = 4, \\ w_1 = 2, & w_2 = 4, & w_3 = 6, & w_4 = 3, & w_5 = 7, & w_6 = 6. \end{matrix}$$

Since the weights w_j are integers, the possible values for x_j are $x_j \in \{0, 1, \ldots, 10\}$, $j = 1(1)n$, and $x_0 = 10$. The table contains the function values $\phi_j(x_{j-1})$ and the actual decision $u_j(x_{j-1})$ for every stage and for every state x_{j-1}. For example, the values of $\phi_6(x_5)$, $\phi_3(2)$, $\phi_3(6)$, and $\phi_3(8)$ are calculated:

$$\phi_6(x_5) = \begin{cases} 0, & \text{if } x_5 < w_6 = 4, \\ c_6 = 6, & \text{otherwise,} \end{cases} \qquad u_6(x_5) = \begin{cases} 0, \text{ if } x_5 < 4, \\ 0, \text{ otherwise.} \end{cases}$$

$\phi_3(2)$: $\quad x_2 = 2 < w_3 = 3$: $\quad \phi_3(2) = \phi_4(2) = 3, \ u_3(2) = 0.$

$\phi_3(6)$: $\quad x_2 > w_3$ and $c_3 + \phi_3(x_2 - w_3) = 6 + 3 < \phi_4(x_2) = 10$: $\quad \phi_3(6) = 10, \ u_3(6) = 0.$

$\phi_3(8)$: $\quad x_2 > w_3$ and $c_3 + \phi_3(x_2 - w_3) = 6 + 9 > \phi_4(x_2) = 10$: $\quad \phi_3(8) = 15, \ u_3(8) = 1.$

The optimal policy is

$$(u_1^*, u_2^*, u_3^*, u_4^*, u_5^*, u_6^*) = (0, 1, 1, 1, 0, 1), \quad \phi_1(10) = 19.$$

	$x_j = 0$	1	2	3	4	5	6	7	8	9	10
$j = 1$											19; 0
2	0; 0	3; 0	4; 1	7; 1	9; 0	10; 1	13; 1	13; 1	15; 0	16; 0	19; 1
3	0; 0	3; 0	3; 0	6; 1	9; 1	9; 0	10; 0	12; 1	15; 1	16; 1	16; 0
4	0; 0	3; 1	3; 1	3; 1	6; 0	9; 1	10; 1	10; 1	10; 1	13; 0	16; 1
5	0; 0	0; 0	0; 0	0; 0	6; 0	7; 1	7; 1	7; 1	7; 1	13; 1	13; 1
6	0; 0	0; 0	0; 0	0; 0	6; 1	6; 1	6; 1	6; 1	6; 1	6; 1	6; 1

19 Numerical Analysis

The most important principles of numerical analysis will be the subject of this chapter. The solution of practical problems usually requires the application of a professional *numerical library* of numerical methods, developed for computers. Some of them will be introduced at the end of Section 19.8.3. The special computer algebra system Mathematica will be discussed with its numerical programs in Chapter 20, p. 1023 and in Section 19.8.4.2, p. 1016. Error propagation and computation errors will be examined in Section 19.8.2, p. 1004.

19.1 Numerical Solution of Non-Linear Equations in a Single Unknown

Every equation with one unknown can be transformed into one of the normal forms:

Zero form: $f(x) = 0.$ (19.1)

Fixed point form: $x = \varphi(x).$ (19.2)

Suppose equations (19.1) and (19.2) can be solved. The solutions are denoted by x^*. To get a first approximation of x^*, we can try to transfom the equation into the form $f_1(x) = f_2(x)$, where the curves of the functions $y = f_1(x)$ and $y = f_2(x)$ are more or less simple to sketch.

■ $f(x) = x^2 - \sin x = 0$. From the shapes of the curves $y = x^2$ and $y = \sin x$ it can be seen that $x_1^* = 0$ and $x_2^* \approx 0.87$ are roots (**Fig. 19.1**).

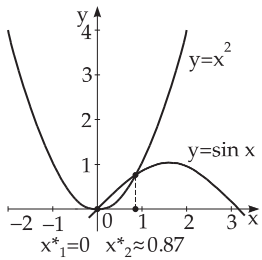


Figure 19.1

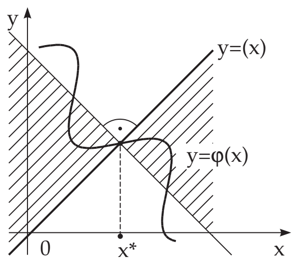


Figure 19.2

19.1.1 Iteration Method

The general idea of iterative methods is that starting with known initial approximations x_k $(k = 0, 1, \ldots, n)$ a sequence of further and better approximations is formed, step by step, hence the solution of the given equation is approached by *iteration*, by a convergent sequence. A sequence is tried to be created with convergence as fast as possible.

19.1.1.1 Ordinary Iteration Method

To solve an equation given in or transformed into the fixed point form $x = \varphi(x)$, the iteration rule

$$x_{n+1} = \varphi(x_n) \quad (n = 0, 1, 2, \ldots;\ x_0 \text{ given}) \tag{19.3}$$

is used which is called the *ordinary iteration method.* It converges to a solution x^* if there is a neighborhood of x^* (**Fig. 19.2**) such that

$$\left|\frac{\varphi(x) - \varphi(x^*)}{x - x^*}\right| \le K < 1 \quad (K \text{ const}) \tag{19.4}$$

I.N. Bronshtein et al., *Handbook of Mathematics*,
DOI 10.1007/978-3-662-46221-8_19

holds, and the initial approximation x_0 is in this neighborhood. If $\varphi(x)$ is differentiable, then the corresponding condition is

$$|\varphi'(x)| \le K < 1. \tag{19.5}$$

The convergence of the ordinary iteration method becomes faster with smaller values of K.

■ $x^2 = \sin x$, i.e., $x_{n+1} = \sqrt{\sin x_n}$.

n	0	1	2	3	4	5
x_n	0.87	0.8742	0.8758	0.8764	0.8766	0.8767
$\sin x_n$	0.7643	0.7670	0.7681	0.7684	0.7686	0.7686

Remark 1: In the case of complex solutions, $x = u + iv$ is substituted. Separating the real and the imaginary part, a system of two equations is obtained for the real unknowns u and v.

Remark 2: The iterative solution of non-linear equation systems can be found in 19.2.2, p. 961.

19.1.1.2 Newton's Method

1. Formula of the Newton Method

To solve an equation given in the form $f(x) = 0$, mostly the *Newton method* is used which has the formula

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \quad (n = 0, 1, 2, \dots\,;\ x_0 \text{ is given}), \tag{19.6}$$

i.e., to get a new approximation x_{n+1}, the values of the function $f(x)$ and its first derivative $f'(x)$ at x_n are needed.

2. Convergence of the Newton Method

The condition

$$f'(x) \ne 0 \tag{19.7a}$$

is necessary for convergence of the Newton method, and the condition

$$\left|\frac{f(x)f''(x)}{f'^2(x)}\right| \le K < 1 \quad (K \text{ const}) \tag{19.7b}$$

is sufficient. The conditions (19.7a,b) must be fulfilled in a neighborhood of x^* such that it contains all the points x_n and x^* itself. If the Newton method is convergent, it converges very fast. It has quadratic convergence, which means that the error of the $(n+1)$-st approximation is less than a constant multiple of the square of the error of the n-th approximation. In the decimal system, this means that after a while the number of exact digits will double step by step.

■ The solution of the equation $f(x) = x^2 - a = 0$, i.e., the calculation of $x = \sqrt{a}$ ($a > 0$ is given), with the Newton method results in the iteration formula

$$x_{n+1} = \frac{1}{2}\left(x_n + \frac{a}{x_n}\right). \tag{19.8}$$

We get for $a = 2$:

n	0	1	2	3
x_n	1.5	1.416 666 6	1.414 215 7	1.414 213 6

3. Geometric Interpretation

The geometric interpretation of the Newton method is represented in **Fig. 19.3**. The basic idea of the Newton method is the local approximation of the curve $y = f(x)$ by its tangent line.

4. Modified Newton Method

If the values of $f'(x_n)$ barely change during the iteration, it can be kept as constant for a while, and the so-called modified Newton method can be used:

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_m)} \quad (m \text{ fixed}, \ m < n). \tag{19.9}$$

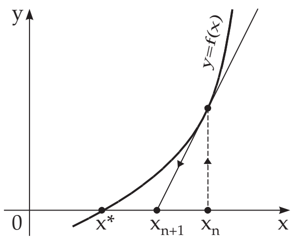


Figure 19.3

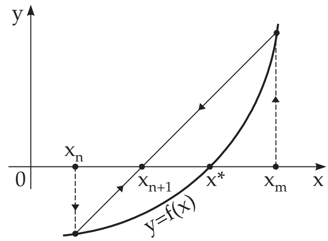


Figure 19.4

The goodness of the convergence is hardly modified by this simplification.

5. Differentiable Functions with Complex Argument

The Newton method also works for differentiable functions with complex arguments.

19.1.1.3 Regula Falsi

1. Formula for Regula Falsi

To solve the equation $f(x) = 0$, the *regula falsi* method has the rule:

$$x_{n+1} = x_n - \frac{x_n - x_m}{f(x_n) - f(x_m)} f(x_n) \quad (n = 1, 2, \ldots\,; \;\; m < n\,; \, x_0, x_1 \text{ are given}). \tag{19.10}$$

Only the function values are calculated. The method follows from the Newton method (19.6) by approximating the derivative $f'(x_n)$ by the finite difference of $f(x)$ between x_n and a previous approximation x_m $(m < n)$.

2. Geometric Interpretation

The geometric interpretation of the regula falsi method is represented in **Fig. 19.4**. The basic idea of the regula falsi method is the local approximation of the curve $y = f(x)$ by a secant line.

3. Convergence

The method (19.10) is convergent if m is chosen so that $f(x_m)$ and $f(x_n)$ always have different signs. If the convergence already seems to be fast enough during the process, it will speed up if one ignores the change of sign, and substitutes $x_m = x_{n-1}$.

■ $f(x) = x^2 - \sin x = 0$.

n	$\Delta x_n = x_n - x_{n-1}$	x_n	$f(x_n)$	$\Delta y_n = f(x_n) - f(x_{n-1})$	$\dfrac{\Delta x_n}{\Delta y_n}$
0		0.9	0.0267		
1	−0.3	0.87	−0.0074	−0.0341	0.8798
2	0.0065	0.8765	−0.000252	0.007148	0.9093
3	0.000229	0.876729	0.000003	0.000255	0.8980
4	−0.000003	0.876726			

If during the process the value of $\Delta x_n/\Delta y_n$ only barely changes, one does not need to recalculate it again and again.

4. Steffensen Method

Applying the regula falsi method with $x_m = x_{n-1}$ for the equation $f(x) = x - \varphi(x) = 0$ the convergence often can be sped up, especially in the case $\varphi'(x) < -1$. This algorithm is known as the *Steffensen method*.

■ To solve the equation $x^2 = \sin x$ with the Steffensen method, the form $f(x) = x - \sqrt{\sin x} = 0$ should be used.

n	$\Delta x_n = x_n - x_{n-1}$	x_n	$f(x_n)$	$\Delta y = f(x_n) - f(x_{n-1})$	$\dfrac{\Delta x_n}{\Delta y_n}$
0		0.9	0.014942		
1	−0.03	0.87	−0.004259	−0.019201	1.562419
2	0.006654	0.876654	−0.000046	0.004213	1.579397
3		0.876727	0.000001		

19.1.2 Solution of Polynomial Equations

Polynomial equations of n-th degree have the form

$$f(x) = p_n(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 = 0. \tag{19.11}$$

For their effective solution efficient methods are needed to calculate the function values and the derivative values of the function $p_n(x)$ and an initial estimate of the positions of the roots.

19.1.2.1 Horner's Scheme

1. Real Arguments

To determine the value of a polynomial $p_n(x)$ of n-th degree at the point $x = x_0$ from its coefficients, first the decomposition

$$p_n(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_2 x^2 + a_1 x + a_0 = (x - x_0)p_{n-1}(x) + p_n(x_0) \tag{19.12}$$

is considered where $p_{n-1}(x)$ is a polynomial of $(n-1)$-st degree:

$$p_{n-1}(x) = a'_{n-1}x^{n-1} + a'_{n-2}x^{n-2} + \cdots + a'_1 x + a'_0. \tag{19.13}$$

The recursion formula

$$a'_{k-1} = x_0 a'_k + a_k, \quad (k = n, n-1, \dots, 0;\ a'_n = 0,\ \ a'_{-1} = p_n(x_0)) \tag{19.14}$$

is obtained by coefficient comparison in (19.12) with respect to x^k. (Note that $a'_{n-1} = a_n$.) In this way, the coefficients a'_k of $p_{n-1}(x)$ and the value $p_n(x_0)$ are determined from the coefficients a_k of $p_n(x)$. Furthermore fewer multiplications are required than by the "traditional" way. By repeating this procedure, a decomposition of the polynomial $p_{n-1}(x)$ with the polynomial $p_{n-2}(x)$ is obtained,

$$p_{n-1}(x) = (x - x_0)p_{n-2}(x) + p_{n-1}(x_0) \tag{19.15}$$

etc., and a sequence of polynomials $p_n(x)$, $p_{n-1}(x)$ $\dots$, $p_1(x)$, $p_0(x)$ is resulted. The calculations of the coefficients and the values of the polynomial is represented in (19.16).

$$\begin{array}{c|llllllll}
 & a_n & a_{n-1} & a_{n-2} & \dots\ a_3 & a_2 & a_1 & a_0 & \\
x_0 & & x_0a'_{n-1} & x_0a'_{n-2} & \dots\ x_0a'_3 & x_0a'_2 & x_0a'_1 & x_0a'_0 & \\
\hline
 & a'_{n-1} & a'_{n-2} & a'_{n-3} & \dots\ a'_2 & a'_1 & a'_0 & \boxed{p_n(x_0)} & \\
x_0 & & x_0a''_{n-2} & x_0a''_{n-3} & \dots\ x_0a''_2 & x_0a''_1 & x_0a''_0 & & \\
\hline
 & a''_{n-2} & a''_{n-3} & a''_{n-4} & \dots\ a''_1 & a''_0 & \boxed{p_{n-1}(x_0)} & & \\
 & & \dots\dots\dots & \dots\dots\dots & \dots\dots & & & & \\
x_0 & & x_0a_0^{(n-1)} & & & & & & \\
\hline
 & a_1^{(n-1)} & \boxed{p_1(x_0)} & & & & & & \\
x_0 & & & & & & & & \\
\hline
 & a_0^{(n)} = p_0(x_0) & & & & & & &
\end{array} \tag{19.16}$$

From scheme (19.16) the value $p_n(x_0)$, and derivatives $p_n^{(k)}(x_0)$ are as:

$$p'_n(x_0) = 1!p_{n-1}(x_0), \ \ p''_n(x_0) = 2!p_{n-2}(x_0), \dots ,\ p_n^{(n)}(x_0) = n!p_0(x_0). \tag{19.17}$$

■ $p_4(x) = x^4 + 2x^3 - 3x^2 - 7$. The substitution value and derivatives of $p_4(x)$ are calculated at $x_0 = 2$ according to (19.16).

$$
\begin{array}{c|ccccc}
 & 1 & 2 & -3 & 0 & -7 \\
2 & & 2 & 8 & 10 & 20 \\
\hline
 & 1 & 4 & 5 & 10 & \boxed{13} \\
2 & & 2 & 12 & 34 & \\
\hline
 & 1 & 6 & 17 & \boxed{44} & \\
2 & & 2 & 16 & & \\
\hline
 & 1 & 8 & \boxed{33} & & \\
2 & & 2 & & & \\
\hline
 & 1 & \boxed{10} & & & \\
2 & & & & & \\
\hline
 & 1 & & & &
\end{array}
$$

We see:
$p_4(2) = 13,$
$p_4'(2) = 44,$
$p_4''(2) = 66,$
$p_4'''(2) = 60,$
$p_4^{(4)}(2) = 24.$

Remarks:

1. The polynomial $p_n(x)$ can be rearranged with respect to the powers of $x - x_0$, e.g., in the example above there is $p_4(x) = (x-2)^4 + 10(x-2)^3 + 33(x-2)^2 + 44(x-2) + 13$.

2. The Horner scheme can also be used for complex coefficients a_k. In this case for every coefficient we have to compute a real and an imaginary column according to (19.16).

2. Complex Arguments

If the coefficients a_k in (19.11) are real, then the calculation of $p_n(x_0)$ for complex values $x_0 = u_0 + \mathrm{i}v_0$ can be made real. In order to show this, $p_n(x)$ is decomposed as follows:

$$
\begin{aligned}
p_n(x) &= a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 \\
&= (x^2 - px - q)(a'_{n-2}x^{n-2} + \cdots + a'_0) + r_1 x + r_0 \quad \text{with}
\end{aligned} \tag{19.18a}
$$

$$
x^2 - px - q = (x - x_0)(x - \overline{x}_0), \text{ i.e., } p = 2u_0,\ q = -({u_0}^2 + {v_0}^2). \tag{19.18b}
$$

Then,

$$
p_n(x_0) = r_1 x + r_0 = (r_1 u_0 + r_0) + \mathrm{i} r_1 v_0. \tag{19.18c}
$$

To find (19.18a) the so-called *two-row Horner scheme* introduced by Collatz can be constructed:

$$
\begin{array}{c|cccccccc}
 & a_n & a_{n-1} & a_{n-2} & \ldots & a_3 & a_2 & a_1 & a_0 \\
q & & & qa'_{n-2} & \ldots & qa'_3 & qa'_2 & qa'_1 & qa'_0 \\
p & & pa'_{n-2} & pa'_{n-3} & \ldots & pa'_2 & pa'_1 & pa'_0 & \\
\hline
 & a'_{n-2} & a'_{n-3} & a'_{n-4} & \ldots & a'_1 & a'_0 & \multicolumn{1}{|c}{r_1} & r_0 \\
 & = a_n & & & & & & &
\end{array} \tag{19.18d}
$$

■ $p_4(x) = x^4 + 2x^3 - 3x^2 - 7$. Calculate the value of p_4 at $x_0 = 2 - \mathrm{i}$, i.e., for $p = 4$ and $q = -5$.

$$
\begin{array}{r|ccccc}
 & 1 & 2 & -3 & 0 & -7 \\
-5 & & & -5 & -30 & -80 \\
4 & & 4 & 24 & 64 & \\
\hline
 & 1 & 6 & 16 & \multicolumn{1}{|c}{34} & -87
\end{array}
$$

It results in:
$p_4(x_0) = 34x_0 - 87 = -19 - 34\mathrm{i}$.

19.1.2.2 Positions of the Roots

1. Real Roots, Sturm Sequence

The *Cartesian rule of signs* gives a first idea of whether the polynomial equation (19.11) has a real root, or not.

a) The number of positive roots is equal to the number of sign changes in the sequence of the cooefficients

$$
a_n,\ a_{n-1}, \ldots,\ a_1,\ a_0 \tag{19.19a}
$$

or it is less by an even number.

b) The number of negative roots is equal to the number of sign changes in the coefficient sequence

$$a_0,\ -a_1,\ a_2, \ldots,\ (-1)^n a_n \tag{19.19b}$$

or it is less by an even number.

■ $p_5(x) = x^5 - 6x^4 + 10x^3 + 13x^2 - 15x - 16$ has 1 or 3 positive roots and 0 or 2 negative roots.

To determine the number of real roots in any given interval (a, b), *Sturm sequences* are used (see 1.6.3.2, **2.**, p. 44).

After computing the function values $y_\nu = p_n(x_\nu)$ at a uniformly distributed set of nodes $x_\nu = x_0 + \nu \cdot h$ (h constant, $\nu = 0, 1, \ldots$) (which can be easily performed by using the Horner scheme) a good guess of the graph of the function and the locations of roots are obtained. If $p_n(c)$ and $p_n(d)$ have different signs, there is at least one real root between c and d.

2. Complex Roots

In order to localize the real or complex roots into a bounded region of the complex plane the following polynomial equation is considered which is a simple consequence of (19.11):

$$f^*(x) = |a_{n-1}|r^{n-1} + |a_{n-2}|r^{n-2} + \cdots + |a_1|r + |a_0| = |a_n|r^n \tag{19.20}$$

and an upper bound r_0 is determined for the positive roots of (19.20), e.g., by systematic repeated trial and error. Then, for all roots x_k^* $(k = 1, 2, \ldots, n)$ of (19.11),

$$|x_k^*| \le r_0. \tag{19.21}$$

■ $f(x) = p_4(x) = x^4 + 4.4x^3 - 20.01x^2 - 50.12x + 29.45 = 0$, $f^*(x) = 4.4r^3 + 20.01r^2 + 50.12r + 29.45 = r^4$. Some trials are

$$\begin{aligned} r = 6:&\quad f^*(6) = 2000.93 > 1296 = r^4,\\ r = 7:&\quad f^*(7) = 2869.98 > 2401 = r^4,\\ r = 8:&\quad f^*(8) = 3963.85 < 4096 = r^4. \end{aligned}$$

From this it follows that $|x_k^*| < 8$ $(k = 1, 2, 3, 4)$. Actually, for the root x_1^* with maximal absolute value $-7 < x_1^* < -6$ holds.

Remark: A special method has been developed in electrotechnics in the so-called *root locus theory* for the determination of the number of complex roots with negative real parts. It is used to examine stability (see [19.11], [19.31]).

19.1.2.3 Numerical Methods

1. General Methods

The methods discussed in Section 19.1.1, p. 949, can be used to find real roots of polynomial equations. The Newton method is well suited for polynomial equations because of its fast convergence, and the fact that the values of $f(x_n)$ and $f'(x_n)$ can be easily computed by using Horner's rule. By assuming that an approximation x_n of the root x^* of a polynomial equation $f(x) = 0$ is sufficiently good, then the correction term $\delta = x^* - x_n$ can be iteratively improved by using the fixed-point equation

$$\delta = -\frac{1}{f'(x_n)}\left[f(x_n) + \frac{1}{2!}f''(x_n)\delta^2 + \cdots\right] = \varphi(\delta). \tag{19.22}$$

2. Special Methods

The *Bairstow method* is well applicable to find root pairs, especially complex conjugate pairs of roots. It starts with finding a quadratic factor of the given polynomial like the Horner scheme (19.18a–d) by determinung the coefficients p and q which make the coefficients of the linear remainder r_0 and r_1 equal to zero (see [19.30], [19.11], [19.31]).

If the computation of the root with largest or smallest absolute value is required, then the *Bernoulli method* is the choice (see [19.19]).

The *Graeffe method* has some historical importance. It gives all roots simultaneously including complex conjugate roots; however the computation costs are tremendous (see [19.11], [19.31]).

19.2 Numerical Solution of Systems of Equations

In several practical problems, there are m conditions for the n unknown quantities $x_i \ (i = 1, 2, \ldots, n)$ in the form of equations:

$$\begin{aligned} F_1(x_1, x_2, \ldots, x_n) &= 0, \\ F_2(x_1, x_2, \ldots, x_n) &= 0, \\ \vdots \qquad\qquad & \ \ \vdots \\ F_m(x_1, x_2, \ldots, x_n) &= 0. \end{aligned} \tag{19.23}$$

The unknowns x_i are to be determined so that they form a solution of the system of equations (19.23). Mostly $m = n$ holds, i.e., the number of unknowns and the number of equations are equal to each other. In the case of $m > n$, (19.23) is called an *over-determined system*; in the case of $m < n$ it is an *under-determined system*.

Over-determined systems usually have no solutions. Then one looks for the "best" solution of (19.23), in the Euclidean metric with the *least squares method*

$$\sum_{i=1}^{m} F_i^2(x_1, x_2, \ldots, x_n) = \min! \tag{19.24}$$

or in other metrics as another extreme value problem. Usually, the values of $n-m$ variables of an under-determined problem can be chosen freely, so the solution of (19.23) depends on $n - m$ parameters. It is called an $(n - m)$-dimensional *manifold of solutions*.

Linear and *non-linear systems of equations* are distinguished, depending on whether the equations are only linear or also non-linear in the unknowns.

19.2.1 Systems of Linear Equations

Consider the linear system of equations

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1, \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2, \\ \vdots \qquad\qquad\qquad & \ \ \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n &= b_n. \end{aligned} \tag{19.25}$$

The system (19.25) can be written in matrix form

$$\mathbf{A}\,\underline{\mathbf{x}} = \underline{\mathbf{b}} \tag{19.26a}$$

with

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & & & \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}, \qquad \underline{\mathbf{b}} = \begin{pmatrix} b_1, \\ b_2, \\ \vdots \\ b_n \end{pmatrix}, \qquad \underline{\mathbf{x}} = \begin{pmatrix} x_1, \\ x_2, \\ \vdots \\ x_n \end{pmatrix}. \tag{19.26b}$$

Suppose the quadratic matrix $\mathbf{A} = (a_{ik}) \ (i, k = 1, 2, \ldots, n)$ is regular, so system (19.25) has a unique solution (see 4.5.2.1, **2.**, p. 309). In the practical solution of (19.25) two types of solution methods are distinguished:

1. Direct Methods are based on elementary transformations, from which the solution can be obtained immediately. These are the pivoting techniques (see 4.5.1.2, p. 307) and the methods given in 19.2.1.1–19.2.1.3.

2. Iteration methods start with a known initial approximation of the solution, and form a sequence of approximations that converges to the solution of (19.25) (see 19.2.1.4, p. 960).

19.2.1.1 Triangular Decomposition of a Matrix

1. Principle of the Gauss Elimination Method

By elementary transformations

1. interchanging rows,
2. multiplying a row by a non-zero number and
3. adding a multiple of a row to another row,
the system $\mathbf{A}\,\underline{\mathbf{x}} = \underline{\mathbf{b}}$ is transformed into the so-called *row echelon form*

$$\mathbf{R}\,\underline{\mathbf{x}} = \underline{\mathbf{c}} \ \text{ with } \ \mathbf{R} = \begin{pmatrix} r_{11} & r_{12} & r_{13} & \dots & r_{1n} \\ & r_{22} & r_{23} & \dots & r_{2n} \\ & & r_{33} & \dots & r_{3n} \\ & 0 & & \ddots & \vdots \\ & & & & r_{nn} \end{pmatrix}. \tag{19.27}$$

Since only equivalent transformations were made, the system of equations $\mathbf{R}\,\underline{\mathbf{x}} = \underline{\mathbf{c}}$ has the same solutions as $\mathbf{A}\,\underline{\mathbf{x}} = \underline{\mathbf{b}}$. From (19.27) it follows:

$$x_n = \frac{c_n}{r_{nn}}, \quad x_i = \frac{1}{r_{ii}}\left(c_i - \sum_{k=i+1}^{n} r_{ik}x_k\right) \quad (i = n-1, n-2, \dots, 1)\,. \tag{19.28}$$

The rule given in (19.28) is called *backward substitution*, since the equations of (19.27) are used in the opposite order as they follow each other.
The transition from $\mathbf{A}$ to $\mathbf{R}$ is made by $n-1$ so-called *elimination steps*, whose procedure is shown by the first step. This step transforms matrix $\mathbf{A}$ into matrix $\mathbf{A}_1$:

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ a_{31} & a_{32} & \dots & a_{3n} \\ \vdots & & & \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix}, \qquad \mathbf{A}_1 = \left(\begin{array}{c|ccc|} a_{11}^{(1)} & a_{12}^{(1)} & \dots & a_{1n}^{(1)} \\ \hline 0 & a_{22}^{(1)} & \dots & a_{2n}^{(1)} \\ 0 & a_{32}^{(1)} & \dots & a_{3n}^{(1)} \\ \vdots & & \vdots & \\ 0 & a_{n2}^{(1)} & \dots & a_{nn}^{(1)} \\ \hline \end{array}\right). \tag{19.29}$$

Then:

1. An $a_{r1} \neq 0$ is chosen (according to (19.33)). If there is none, stop: $\mathbf{A}$ is singular. Otherwise a_{r1} is called the *pivot*.

2. The first and the r-th row of $\mathbf{A}$ are interchanged. The result is $\overline{\mathbf{A}}$.

3. The l_{i1} $(i = 2, 3, \dots, n)$ multiple of the first row is subtracted from the i-th row of the matrix $\overline{\mathbf{A}}$. The result is the matrix $\mathbf{A}_1$ and analogously the new right-hand side $\underline{\mathbf{b}}_1$ with the elements

$$a_{ik}^{(1)} = \overline{a}_{ik} - l_{i1}\overline{a}_{1k} \ \text{ with } \ l_{i1} = \frac{\overline{a}_{i1}}{\overline{a}_{11}},$$

$$b_i^{(1)} = \overline{b}_i - l_{i1}\overline{b}_1 \quad (i, k = 2, 3, \dots, n). \tag{19.30}$$

The framed submatrix in $\mathbf{A}_1$ (see (19.29)) is of type $(n-1, n-1)$ and it will be handled analogously to $\mathbf{A}$, etc. This method is called the *Gaussian elimination method* or the *Gauss algorithm* (see 4.5.2.4, p. 312).

2. Triangular Decomposition

The result of the Gauss elimination method can be formulated as follows: To every regular matrix $\mathbf{A}$ there exists a so-called *triangular decomposition* or *LU factorization* of the form

$$\mathbf{P}\,\mathbf{A} = \mathbf{L}\,\mathbf{R} \tag{19.31}$$

with

$$\mathbf{R} = \begin{pmatrix} r_{11} & r_{12} & r_{13} & \dots & r_{1n} \\ & r_{22} & r_{23} & \dots & r_{2n} \\ & & r_{33} & \dots & r_{3n} \\ & 0 & & \ddots & \vdots \\ & & & & r_{nn} \end{pmatrix}, \quad \mathbf{L} = \begin{pmatrix} 1 & & & & \\ l_{21} & 1 & & 0 & \\ l_{31} & l_{32} & 1 & & \\ \vdots & \vdots & & \ddots & \\ l_{n1} & l_{n2} & \dots & l_{n,n-1} & 1 \end{pmatrix}. \tag{19.32}$$

Here $\mathbf{R}$ is called an *upper triangular matrix*, $\mathbf{L}$ is a *lower triangular matrix* and $\mathbf{P}$ is a so-called *permutation matrix*. A permutation matrix is a quadratic matrix which has exactly one 1 in every row and every column, and the other elements are zeros. The multiplication $\mathbf{P}\,\mathbf{A}$ results in row interchanges in $\mathbf{A}$, which comes from the choices of the pivot elements during the elimination procedure.

■ The Gauss elimination method should be used for the system $\begin{pmatrix} 3 & 1 & 6 \\ 2 & 1 & 3 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 2 \\ 7 \\ 4 \end{pmatrix}$. In schematic form, where the coefficient matrix and the vector from the right-hand side are written next to each other (into the so-called *extended coefficient matrix*), the calculations are:

$$(\mathbf{A}, \underline{\mathbf{b}}) = \left(\begin{array}{ccc|c} \boxed{3} & 1 & 6 & 2 \\ 2 & 1 & 3 & 7 \\ 1 & 1 & 1 & 4 \end{array}\right) \Rightarrow \left(\begin{array}{ccc|c} 3 & 1 & 6 & 2 \\ 2/3 & 1/3 & -1 & 17/3 \\ 1/3 & \boxed{2/3} & -1 & 10/3 \end{array}\right) \Rightarrow \left(\begin{array}{ccc|c} 3 & 1 & 6 & 2 \\ 1/3 & 2/3 & -1 & 10/3 \\ 2/3 & 1/2 & \boxed{-1/2} & 4 \end{array}\right), \text{ i.e.,}$$

$$\mathbf{P} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \Rightarrow \mathbf{P}\,\mathbf{A} = \begin{pmatrix} 3 & 1 & 6 \\ 1 & 1 & 1 \\ 2 & 1 & 3 \end{pmatrix}, \mathbf{L} = \begin{pmatrix} 1 & 0 & 0 \\ 1/3 & 1 & 0 \\ 2/3 & 1/2 & 1 \end{pmatrix}, \mathbf{R} = \begin{pmatrix} 3 & 1 & 6 \\ 0 & 2/3 & -1 \\ 0 & 0 & -1/2 \end{pmatrix}.$$

In the extended coefficient matrices, the matrices $\mathbf{A}$, $\mathbf{A}_1$ and $\mathbf{A}_2$, and also the pivots are shown in boxes. Solution: $x_3 = -8$, $x_2 = -7$, $x_1 = 19$.

3. Application of Triangular Decomposition

With the help of triangular decomposition, the solution of the linear system of equations $\mathbf{A}\,\underline{\mathbf{x}} = \underline{\mathbf{b}}$ can be described in three steps:

1. $\mathbf{P}\,\mathbf{A} = \mathbf{L}\,\mathbf{R}$: Determination of the triangular decomposition and substitution $\mathbf{R}\,\underline{\mathbf{x}} = \underline{\mathbf{c}}$.

2. $\mathbf{L}\,\underline{\mathbf{c}} = \mathbf{P}\,\underline{\mathbf{b}}$: Determination of the auxiliary vector $\underline{\mathbf{c}}$ by forward substitution.

3. $\mathbf{R}\,\underline{\mathbf{x}} = \underline{\mathbf{c}}$: Determination of the solution $\underline{\mathbf{x}}$ by backward substitution.

If the solution of a system of linear equations is handled by the expanded coefficient matrix $(\mathbf{A}, \underline{\mathbf{b}})$, as in the above example, by the Gauss elimination method, then the lower triangular matrix $\mathbf{L}$ is not needed explicitly. This can be especially useful if several systems of linear equations are to be solved after each other with the same coefficient matrix, with different right-hand sides.

4. Choice of the Pivot Elements

Theoretically, every non-zero element $a_{i1}^{(k-1)}$ of the first column of the matrix $\mathbf{A}_{k-1}$ could be used as a pivot element at the k-th elimination step. In order to improve the accuracy of solution (to decrease the accumulated rounding errors of the operations), the following strategies are recommended.

1. Diagonal Strategy The diagonal elements are chosen successively as pivot elements if possible, i.e., there is no row interchange. This kind of choice of the pivot element makes sense if the absolute value of the elements of the main diagonal are fairly large compared to the others in the same row.

2. Column Pivoting To perform the k-th elimination step, such row index r is chosen for which:

$$|a_{rk}^{(k-1)}| = \max_{i \geq k} |a_{ik}^{(k-1)}|. \tag{19.33}$$

If $r \neq k$, then the r-th and the k-th rows will be interchanged. It can be proven that this strategy makes the accumulated rounding errors smaller.

19.2.1.2 Cholesky's Method for a Symmetric Coefficient Matrix

In several cases, the coefficient matrix $\mathbf{A}$ in (19.26a) is not only symmetric, but also *positive definite*, i.e., for the corresponding *quadratic form* $Q(\underline{\mathbf{x}})$ holds

$$Q(\underline{\mathbf{x}}) = \underline{\mathbf{x}}^{\mathrm{T}}\mathbf{A}\,\underline{\mathbf{x}} = \sum_{i=1}^{n}\sum_{k=1}^{n} a_{ik}x_i x_k > 0 \tag{19.34}$$

for every $\underline{\mathbf{x}} \in \mathbb{R}^n$, $\underline{\mathbf{x}} \neq \underline{\mathbf{0}}$. Since for every symmetric positive definite matrix $\mathbf{A}$ there exists a unique triangular decomposition

$$\mathbf{A} = \mathbf{L}\,\mathbf{L}^{\mathrm{T}} \tag{19.35}$$

with

$$\mathbf{L} = \begin{pmatrix} l_{11} & & & & \\ l_{21} & l_{22} & & 0 & \\ l_{31} & l_{32} & l_{33} & & \\ \vdots & & & \ddots & \\ l_{n1} & l_{n2} & l_{n3} & \dots & l_{nn} \end{pmatrix}, \tag{19.36a}$$

$$l_{kk} = \sqrt{a_{kk}^{(k-1)}}, \quad l_{ik} = \frac{a_{ik}^{(k-1)}}{l_{kk}} \quad (i = k, k+1, \dots, n)\,; \tag{19.36b}$$

$$a_{ij}^{(k)} = a_{ij}^{(k-1)} - l_{ik}l_{jk} \quad (i, j = k+1, k+2, \dots, n), \tag{19.36c}$$

the solution of the corresponding linear system of equations $\mathbf{A}\,\underline{\mathbf{x}} = \underline{\mathbf{b}}$ can be determined by the *Cholesky method* by the following steps:

1. $\mathbf{A} = \mathbf{L}\,\mathbf{L}^{\mathrm{T}}$: Determination of the so-called *Cholesky decomposition* and substitution $\mathbf{L}^{\mathrm{T}}\underline{\mathbf{x}} = \underline{\mathbf{c}}$.

2. $\mathbf{L}\,\underline{\mathbf{c}} = \underline{\mathbf{b}}$: Determination of the auxiliary vector $\underline{\mathbf{c}}$ by forward substitution.

3. $\mathbf{L}^{\mathrm{T}}\underline{\mathbf{x}} = \underline{\mathbf{c}}$: Determination of the solution $\underline{\mathbf{x}}$ by backward substitution.

For large values of n the computation cost of the Cholesky method is approximately half of that of the LU decomposition given in (19.31), p. 956.

19.2.1.3 Orthogonalization Method

1. Linear Fitting Problem

Suppose an *over-determined linear system of equations*

$$\sum_{k=1}^{n} a_{ik}x_k = b_i \quad (i = 1, 2, \dots, m;\ m > n), \tag{19.37}$$

is given in matrix form

$$\mathbf{A}\underline{\mathbf{x}} = \underline{\mathbf{b}}. \tag{19.38}$$

Suppose the coefficient matrix $\mathbf{A} = (a_{ik})$ with size $(m \times n)$ has full rank n, i.e., its columns are linearly independent. Since an over-determined linear system of equations usually has no solution, instead of (19.37) the so-called *error equations* are considered

$$r_i = \sum_{k=1}^{n} a_{ik}x_k - b_i \quad (i = 1, 2, \dots, m;\ m > n) \tag{19.39}$$

with *residues* r_i, and the sum of their squares should be minimalized:

$$\sum_{i=1}^{m} {r_i}^2 = \sum_{i=1}^{m}\left[\sum_{k=1}^{n} a_{ik}x_k - b_i\right]^2 = F(x_1, x_2, \dots, x_n) = \min! \tag{19.40}$$

The problem (19.40) is called a *linear fitting problem* or a *linear least squares problem* (see also 6.2.5.5, p. 456). The necessary condition for the relative minimum of the *sum of residual squares* $F(x_1, x_2, \ldots, x_n)$ is

$$\frac{\partial F}{\partial x_k} = 0 \quad (k = 1, 2, \ldots, n) \tag{19.41}$$

and it leads to the linear system of equations

$$\mathbf{A}^{\mathrm{T}}\mathbf{A}\underline{\mathbf{x}} = \mathbf{A}^{\mathrm{T}}\underline{\mathbf{b}}. \tag{19.42}$$

The transition from (19.38) to (19.42) is called a *Gauss transformation*, since the system (19.42) arises by applying the *Gaussian least squares method* (see 6.2.5.5, p. 456) for (19.38). Since $\mathbf{A}$ is supposed to be of full rank, $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ is a positive definite matrix of size $(n \times n)$, and the so-called *normal equations* (19.42) can be solved numerically by the Cholesky method (see 19.2.1.2, p. 958).
One can have numerical difficulties with the solution of the normal equations (19.42) if the *condition number* (see [19.24]) of the matrix $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ is too large. The solution $\underline{\mathbf{x}}$ can then have a large relative error. Because of this problem, it is better to use the orthogonalization method for solving numerically linear fitting problems.

2. Orthogonalization Method

The following facts are the basis of the following orthogonalization method for solving a linear least squares problem (19.40):

1. The length of a vector does not change during an orthogonal transformation, i.e., the vectors $\underline{\mathbf{x}}$ and $\underline{\tilde{\mathbf{x}}} = \mathbf{Q}_0\underline{\mathbf{x}}$ with

$$\mathbf{Q}_0^{\mathrm{T}}\mathbf{Q}_0 = \mathbf{E} \tag{19.43}$$

have the same length.

2. For every matrix $\mathbf{A}$ of size (m, n) with maximal rank n $(n < m)$ there exists an orthogonal matrix $\mathbf{Q}$ of size (m, m) such that

$$\mathbf{A} = \mathbf{Q}\hat{\mathbf{R}} \quad (19.44) \qquad \text{with} \quad \mathbf{Q}^{\mathrm{T}}\mathbf{Q} = \mathbf{E} \quad \text{and} \quad \hat{\mathbf{R}} = \begin{pmatrix} \mathbf{R} \\ \mathbf{O} \end{pmatrix} = \begin{pmatrix} r_{11} & r_{12} & \ldots & r_{1n} \\ & r_{22} & \ldots & r_{2n} \\ & & \ddots & \vdots \\ & & & r_{nn} \\ \hline & & \mathbf{O} & \end{pmatrix}. \tag{19.45}$$

Here $\mathbf{R}$ is an upper triangular matrix of size (n, n), and $\mathbf{O}$ is a zero matrix of size $(m - n, n)$.
The factored form (19.44) of matrix $\mathbf{A}$ is called the *QR decomposition*. So, the error equations (19.39) can be transformed into the equivalent system

$$\begin{array}{rl}
r_{11}x_1 + r_{12}x_2 + \ldots + r_{1n}x_n - \hat{b}_1 & = \hat{r}_1, \\
r_{22}x_2 + \ldots + r_{2n}x_n - \hat{b}_2 & = \hat{r}_2, \\
\ddots \quad \vdots \quad \vdots & = \vdots \\
r_{nn}x_n - \hat{b}_n & = \hat{r}_n, \\
-\hat{b}_{n+1} & = \hat{r}_{n+1}, \\
\vdots & \quad \vdots \\
-\hat{b}_m & = \hat{r}_m
\end{array} \tag{19.46}$$

without changing the sum of the squares of the residuals. From (19.46) it follows that the sum of the squares is minimal for $\hat{r}_1 = \hat{r}_2 = \cdots = \hat{r}_n = 0$ and the minimum value is equal to the sum of the squares of $\hat{r}_{n+1}$ to $\hat{r}_m$. The required solution $\underline{\mathbf{x}}$ can be got by backward substitution

$$\mathbf{R}\underline{\mathbf{x}} = \underline{\hat{\mathbf{b}}}_0, \tag{19.47}$$

where $\underline{\hat{\mathbf{b}}}_0$ is the vector with components $\hat{b}_1, \hat{b}_2, \ldots, \hat{b}_n$ obtained from (19.46).
There are two methods most often used for a stepwise transition of (19.39) into (19.46):

1. Givens transformation,

2. Householder transformation.

The first one results in the QR decomposition of matrix $\mathbf{A}$ by *rotations*, the other one by *reflections*. The numerical implementations can be found in [19.23].

Practical problems in linear least squares approximations are solved mostly by the Householder transformation, where the frequently occurring special *band structure* of the coefficient matrix $\mathbf{A}$ can be used.

19.2.1.4 Iteration Methods

1. Jacobi Method

Suppose in the coefficient matrix of the linear system of equations (19.25) every diagonal element a_{ii} $(i = 1, 2, \ldots, n)$ is different from zero. Then the i-th row can be solved for the unknown x_i, and it immediately results the following iteration rule, where μ is the iteration index:

$$x_i^{(\mu+1)} = \frac{b_i}{a_{ii}} - \sum_{\substack{k=1 \\ (k\neq i)}}^{n} \frac{a_{ik}}{a_{ii}} x_k^{(\mu)} \quad (i = 1, 2, \ldots, n) \tag{19.48}$$

$$(\mu = 0, 1, 2, \ldots;\ x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)} \text{ are given initial values}).$$

Formula (19.48) is called the *Jacobi method*. Every component of the new vector $\underline{\mathbf{x}}^{(\mu+1)}$ is calculated from the components of $\underline{\mathbf{x}}^{(\mu)}$. If at least one of the conditions

$$\max_k \sum_{\substack{i=1 \\ (i\neq k)}}^{n} \left|\frac{a_{ik}}{a_{ii}}\right| < 1 \qquad \textit{column sum criterion} \tag{19.49}$$

or

$$\max_i \sum_{\substack{k=1 \\ (k\neq i)}}^{n} \left|\frac{a_{ik}}{a_{ii}}\right| < 1 \qquad \textit{row sum criterion} \tag{19.50}$$

holds, then the Jacobi method is convergent for any initial vector $\underline{\mathbf{x}}^{(0)}$.

2. Gauss-Seidel Method

If the first component $x_1^{(\mu+1)}$ is calculated by the Jacobi method, then this value can be used in the calculation of $x_2^{(\mu+1)}$. While proceeding similarly in the calculation of the further components, the following iteration formula is obtained:

$$x_i^{(\mu+1)} = \frac{b_i}{a_{ii}} - \sum_{k=1}^{i-1} \frac{a_{ik}}{a_{ii}} x_k^{(\mu+1)} - \sum_{k=i+1}^{n} \frac{a_{ik}}{a_{ii}} x_k^{(\mu)} \tag{19.51}$$

$$(i = 1, 2, \ldots, n;\ x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)} \text{ given initial value}; \ \mu = 0, 1, 2, \ldots).$$

Formula (19.51) is called the *Gauss-Seidel method*. The Gauss-Seidel method usually converges faster than the Jacobi method, but its convergence criterion is more complicated.

■
$$\begin{aligned}
10x_1 - 3x_2 - 4x_3 + 2x_4 &= 14, \\
-3x_1 + 26x_2 + 5x_3 - x_4 &= 22, \\
-4x_1 + 5x_2 + 16x_3 + 5x_4 &= 17, \\
2x_1 + 3x_2 - 4x_3 - 12x_4 &= -20.
\end{aligned}$$

The corresponding iteration formula according to (19.51) is:

$$x_1^{(\mu+1)} = \frac{1}{10}\left(14 + 3x_2^{(\mu)} + 4x_3^{(\mu)} - 2x_4^{(\mu)}\right),$$
$$x_2^{(\mu+1)} = \frac{1}{26}\left(22 + 3x_1^{(\mu+1)} - 5x_3^{(\mu)} + x_4^{(\mu)}\right),$$
$$x_3^{(\mu+1)} = \frac{1}{16}\left(17 + 4x_1^{(\mu+1)} - 5x_2^{(\mu+1)} - 5x_4^{(\mu)}\right),$$
$$x_4^{(\mu+1)} = \frac{1}{12}\left(-20 + 2x_1^{(\mu+1)} + 3x_2^{(\mu+1)} - 4x_3^{(\mu+1)}\right).$$

Some approximations and the solution are given here:

$\underline{\mathbf{x}}^{(0)}$	$\underline{\mathbf{x}}^{(1)}$	$\underline{\mathbf{x}}^{(4)}$	$\underline{\mathbf{x}}^{(5)}$	$\underline{\mathbf{x}}$
0	1.4	1.5053	1.5012	1.5
0	1.0077	0.9946	0.9989	1
0	1.0976	0.5059	0.5014	0.5
0	1.7861	1.9976	1.9995	2

3. Relaxation Method

The iteration formula of the Gauss-Seidel method (19.51) can be written in the so-called *correction form*

$$x_i^{(\mu+1)} = x_i^{(\mu)} + \left(\frac{b_i}{a_{ii}} - \sum_{k=1}^{i-1}\frac{a_{ik}}{a_{ii}}x_k^{(\mu+1)} - \sum_{k=i}^{n}\frac{a_{ik}}{a_{ii}}x_k^{(\mu)}\right), \quad \text{i.e.,}$$

$$x_i^{(\mu+1)} = x_i^{(\mu)} + d_i^{(\mu)} \quad (i = 1, 2, \ldots, n;\ \mu = 0, 1, 2, \ldots). \tag{19.52}$$

By an appropriate choice of a *relaxation parameter* ω and rewriting (19.52) in the form

$$x_i^{(\mu+1)} = x_i^{(\mu)} + \omega d_i^{(\mu)} \quad (i = 1, 2, \ldots, n;\ \mu = 0, 1, 2, \ldots), \tag{19.53}$$

one can try to improve the speed of convergence. It can be shown that convergence is possible only for

$$0 < \omega < 2. \tag{19.54}$$

For $\omega = 1$ we retrieve the Gauss-Seidel method. In the case of $\omega > 1$, which is called over-relaxation, the corresponding iteration method is called the *SOR method* (**s**uccessive **o**ver**r**elaxation). The determination of an optimal relaxation parameter is possible only for some special types of matrices.

Iterative methods are applied to solve linear systems of equations in the first place when the main diagonal elements a_{ii} of the coefficient matrix have an absolute value much larger than the other elements a_{ik} $(i \neq k)$ (in the same row or column), or when the rows of the system of equations can be rearranged in a certain way to get such a form.

19.2.2 System of Non-Linear Equations

Suppose the system of n non-linear equations

$$F_i(x_1, x_2, \ldots, x_n) = 0 \quad (i = 1, 2, \ldots, n) \tag{19.55}$$

for the n unknowns $x_1, x_2, \ldots, x_n$ has a solution. Usually, a numerical solution can be given only by an iteration method.

19.2.2.1 Ordinary Iteration Method

The ordinary iteration method can be used if the equations (19.55) can be transformed into a fixed-point form

$$x_i = f_i(x_1, x_2, \ldots, x_n) \quad (i = 1, 2, \ldots, n). \tag{19.56}$$

Then, starting from estimated approximations $x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)}$, the improved values are obtained either by

1. iteration with simultaneous steps

$$x_i^{(\mu+1)} = f_i\left(x_1^{(\mu)}, x_2^{(\mu)}, \ldots, x_n^{(\mu)}\right) \quad (i = 1, 2, \ldots, n;\ \mu = 0, 1, 2, \ldots) \tag{19.57}$$

or by

2. iteration with sequential steps

$$x_i^{(\mu+1)} = f_i\left(x_1^{(\mu+1)}, \ldots, x_{i-1}^{(\mu+1)}, x_i^{(\mu)}, x_{i+1}^{(\mu)}, \ldots, x_n^{(\mu)}\right) \quad (i = 1, 2, \ldots, n;\ \mu = 0, 1, 2, \ldots). \tag{19.58}$$

It is of crucial importance for the convergence of this method that in the neighborhood of the solution the functions f_i should depend only weakly on the unknowns, i.e., if f_i are differentiable, the absolute values of the partial derivatives must be rather small. We get as a *convergence condition*

$$K < 1 \text{ with } K = \max_i \left(\sum_{k=1}^{n} \max \left| \frac{\partial f_i}{\partial x_k} \right| \right). \tag{19.59}$$

With this quantity K, the *error estimation* is the following:

$$\max_i \left| x_i^{(\mu+1)} - x_i \right| \le \frac{K}{1-K} \max_i \left| x_i^{(\mu+1)} - x_i^{(\mu)} \right|. \tag{19.60}$$

Here, x_i is the component of the required solution, $x_i^{(\mu)}$ and $x_i^{(\mu+1)}$ are the corresponding μ-th and $(\mu+1)$-th approximations.

19.2.2.2 Newton's Method

The Newton method is used for the problem given in the form (19.55). After finding the initial approximation values $x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)}$, the functions F_i are expanded in Taylor form as functions of n independent variables $x_1, x_2, \ldots, x_n$ (see p. 471). Terminating the expansion after the linear terms, from (19.55) a linear system of equations is obtained, and iterative improvements can be got by the following formula:

$$F_i\left(x_1^{(\mu)}, x_2^{(\mu)}, \ldots, x_n^{(\mu)}\right) + \sum_{k=1}^{n} \frac{\partial F_i}{\partial x_k}\left(x_1^{(\mu)}, \ldots, x_n^{(\mu)}\right)\left(x_k^{(\mu+1)} - x_k^{(\mu)}\right) = 0 \tag{19.61}$$

$$(i = 1, 2, \ldots, n;\ \mu = 0, 1, 2, \ldots).$$

The coefficient matrix of the linear system of equations (19.61), which should be solved in every iteration step, is

$$\mathbf{F}'(\underline{\mathbf{x}}^{(\mu)}) = \left(\frac{\partial F_i}{\partial x_k}\left(x_1^{(\mu)}, x_2^{(\mu)}, \ldots, x_n^{(\mu)}\right) \right) \quad (i, k = 1, 2, \ldots, n) \tag{19.62}$$

and it is called the *Jacobian matrix*. If the Jacobian matrix is invertible in the neighborhood of the solution, the Newton method is locally quadratically convergent, i.e., its convergence essentially depends on how good the initial approximations are. If $x_k^{(\mu+1)} - x_k^{(\mu)} = d_k^{(\mu)}$ are substituted in (19.61), then the Newton method can be written in the correction form

$$x_k^{(\mu+1)} = x_k^{(\mu)} + d_k^{(\mu)} \quad (i = 1, 2, \ldots, n;\ \mu = 0, 1, 2, \ldots). \tag{19.63}$$

To reduce the sensitivity to the initial values, analogously to the relaxation method, a so-called *damping* or *step length parameter* γ can be introduced (*damping method*):

$$x_k^{(\mu+1)} = x_k^{(\mu)} + \gamma d_k^{(\mu)} \quad (i = 1, 2, \ldots, n;\ \mu = 0, 1, 2, \ldots;\ \gamma > 0). \tag{19.64}$$

Methods to determine γ can be found in [19.24].

19.2.2.3 Derivative-Free Gauss-Newton Method

To solve the least squares problem (19.24), one proceeds iteratively in the non-linear case as follows:

1. Starting from a suitable initial approximation $x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)}$, the non-linear functions $F_i(x_1, x_2, \ldots, x_n)$ $(i = 1, 2, \ldots, m; m > n)$ are approximated as in the Newton method (see (19.61)) by linear approximations $\tilde{F}_i(x_1, x_2, \ldots, x_n)$, which are calculated in every iteration step according to

$$\tilde{F}_i(x_1, \ldots, x_n) = F_i\left(x_1^{(\mu)}, x_2^{(\mu)}, \ldots, x_n^{(\mu)}\right) + \sum_{k=1}^{n} \frac{\partial F_i}{\partial x_k}\left(x_1^{(\mu)}, \ldots, x_n^{(\mu)}\right)\left(x_k - x_k^{(\mu)}\right)$$

$$(i = 1, 2, \ldots, m;\ \mu = 0, 1, 2, \ldots). \tag{19.65}$$

2. $d_k^{(\mu)} = x_k - x_k^{(\mu)}$ are substituted in (19.65) and the corrections $d_k^{(\mu)}$ are determined by using the Gaussian least squares method, i.e., by the solution of the linear least squares problem

$$\sum_{i=1}^{m} \tilde{F}_i^2(x_1, \ldots, x_n) = \min, \tag{19.66}$$

e.g., with the help of the normal equations (see (19.42)), or the Householder method (see 19.6.2.2, p. 985).

3. The approximations for the required solution are given by the formulas

$$x_k^{(\mu+1)} = x_k^{(\mu)} + d_k^{(\mu)} \quad \text{or} \tag{19.67a}$$

$$x_k^{(\mu+1)} = x_k^{(\mu)} + \gamma d_k^{(\mu)} \quad (k = 1, 2, \ldots, n), \tag{19.67b}$$

where γ $(\gamma > 0)$ is a step length parameter similar to the Newton method.

By repeating steps 2 and 3 with $x_k^{(\mu+1)}$ instead of $x_k^{(\mu)}$ one gets the *Gauss-Newton method*. It results in a sequence of approximation values, whose convergence strongly depends on the accuracy of the initial approximation. The sum of the error squares can be reduced by introducing a length parameter γ.

If the evaluation of the partial derivatives $\frac{\partial F_i}{\partial x_k}\left(x_1^{(\mu)}, \ldots, x_n^{(\mu)}\right)$ $(i = 1, 2, \ldots, m;\ k = 1, 2, \ldots, n)$ requires too much work, the partial derivatives can be approximated by difference quotients :

$$\frac{\partial F_i}{\partial x_k}\left(x_1^{(\mu)}, \ldots, x_k^{(\mu)}, \ldots, x_n^{(\mu)}\right) \approx \frac{1}{h_k^{(\mu)}}\Big[F_i\left(x_1^{(\mu)}, \ldots, x_{k-1}^{(\mu)}, x_k^{(\mu)} + h_k^{(\mu)}, x_{k+1}^{(\mu)}, \ldots, x_n^{(\mu)}\right)$$

$$-F_i\left(x_1^{(\mu)}, \ldots, x_k^{(\mu)}, \ldots, x_n^{(\mu)}\right)\Big] \quad (i = 1, 2, \ldots, m;\ k = 1, 2, \ldots, n;\ \mu = 0, 1, 2, \ldots). \tag{19.68}$$

The so-called *discretization step sizes* $h_k^{(\mu)}$ may depend on the iteration steps and the values of the variables.

If the approximations (19.68) are used, then only function values F_i are to be calculated while performing the Gauss-Newton method, i.e., the method is *derivative free*.

19.3 Numerical Integration

19.3.1 General Quadrature Formulas

The numerical evaluation of the definite integral

$$I(f) = \int_a^b f(x)\, dx \tag{19.69}$$

must be done only approximately if the integrand $f(x)$ cannot be integrated by elementary calculus, or it is too complicated, or when the function is known only at certain points x_ν, at the so-called *interpolation nodes* from the integration interval $[a, b]$. The so-called *quadrature formulas* are used for the approximate calculation of (19.69). They have the general form

$$Q(f) = \sum_{\nu=0}^{n} c_{0\nu} y_\nu + \sum_{\nu=0}^{n} c_{1\nu} y_\nu^{(1)} + \cdots + \sum_{\nu=0}^{n} c_{p\nu} y_\nu^{(p)} \tag{19.70}$$

with $y_\nu^{(\mu)} = f^{(\mu)}(x_\nu)$ $(\mu = 1, 2, \ldots, p;\ \nu = 1, 2, \ldots, n)$, $y_\nu = f(x_\nu)$, and constant values of $c_{\mu\nu}$. Obviously,

$$I(f) = Q(f) + R, \tag{19.71}$$

where R is the error of the quadrature formula. In the application of quadrature formulas it is supposed that the required values of the integrand $f(x)$ and its derivatives at the interpolation nodes are known

as numerical values. Formulas using only the values of the function are called *mean value formulas*; formulas using also the derivatives are called *Hermite quadrature formulas*.

19.3.2 Interpolation Quadratures

The following formulas represent so-called *interpolation quadratures*. Here, the integrand $f(x)$ is interpolated at certain interpolation nodes (possibly the least number of them) by a polynomial $p(x)$ of corresponding degree, and the integral of $f(x)$ is replaced by that of $p(x)$. The formula for the integral over the entire interval is given by summation. Here the formulas are given for the most practical cases. The interpolation nodes are *equidistant*:

$$x_\nu = x_0 + \nu h \quad (\nu = 0, 1, 2, \dots, n), \quad x_0 = a, \ \ x_n = b, \ \ h = \frac{b-a}{n}. \tag{19.72}$$

An upper bound for the magnitude of the error $|R|$ is given for every quadrature formula. Here, M_μ means an upper bound of $|f^{(\mu)}(x)|$ on the entire domain.

19.3.2.1 Rectangular Formula

In the interval $[x_0, x_0+h]$, $f(x)$ is replaced by the constant function $y = y_0 = f(x_0)$, which interpolates $f(x)$ at the interpolation node x_0, which is the left endpoint of the integration interval. In this way the *simple rectangular formula* is obtained:

$$\int\limits_{x_0}^{x_0+h} f(x)\,dx \approx h \cdot y_0, \quad |R| \le \frac{h^2}{2} M_1. \tag{19.73a}$$

The *left-sided rectangular formula* by summation is

$$\int\limits_a^b f(x)\,dx \approx h(y_0 + y_1 + y_2 + \cdots + y_{n-1}), \quad |R| \le \frac{(b-a)h}{2} M_1. \tag{19.73b}$$

M_1 denotes an upper bound of $|f'(x)|$ on the entire domain of integration.
One gets analogously the *right-sided rectangular sum*, if one replaces y_0 by y_1 in (19.73a). The formula is

$$\int\limits_a^b f(x)\,dx \approx h(y_1 + y_2 + \cdots + y_n), \quad |R| \le \frac{(b-a)h}{2} M_1. \tag{19.74}$$

19.3.2.2 Trapezoidal Formula

$f(x)$ is replaced by a polynomial of first degree in the interval $[x_0, x_0 + h]$, which interpolates $f(x)$ at the interpolation nodes x_0 and $x_1 = x_0 + h$. The approximation is

$$\int\limits_{x_0}^{x_0+h} f(x)\,dx \approx \frac{h}{2}(y_0 + y_1), \quad |R| \le \frac{h^3}{12} M_2. \tag{19.75}$$

The so-called *trapezoidal formula* can be obtained by summation:

$$\int\limits_a^b f(x)\,dx \approx h\left(\frac{y_0}{2} + y_1 + y_2 + \cdots + y_{n-1} + \frac{y_n}{2}\right), \quad |R| \le \frac{(b-a)h^2}{12} M_2. \tag{19.76}$$

M_2 denotes an upper bound of $|f''(x)|$ on the entire integration domain. The error of the trapezoidal formula is proportional to h^2, i.e., the trapezoidal sum has an error of order 2. It follows that it converges

to the definite integral for $h \to 0$ (hence, $n \to \infty$), if rounding errors are not considered.

19.3.2.3 Simpson's Formula

$f(x)$ is replaced by a polynomial of second degree in the interval $[x_0, x_0 + 2h]$, which interpolates $f(x)$ at the interpolation nodes x_0, $x_1 = x_0 + h$ and $x_2 = x_0 + 2h$:

$$\int\limits_{x_0}^{x_0+2h} f(x)\,dx \approx \frac{h}{3}(y_0 + 4y_1 + y_2), \quad |R| \le \frac{h^5}{90} M_4. \tag{19.77}$$

n must be an even number for a complete Simpson formula. The approximation is

$$\int\limits_a^b f(x)\,dx \approx \frac{h}{3}(y_0 + 4y_1 + 2y_2 + 4y_3 + \cdots + 2y_{n-2} + 4y_{n-1} + y_n), \tag{19.78}$$

$$|R| \le \frac{(b-a)h^4}{180} M_4.$$

M_4 is an upper bound for $|f^{(4)}(x)|$ on the entire integration domain. The Simpson formula has an error of order 4 and it is exact for polynomials up to third degree.

19.3.2.4 Hermite's Trapezoidal Formula

$f(x)$ is replaced by a polynomial of third degree in the interval $[x_0, x_0 + h]$, which interpolates $f(x)$ and $f'(x)$ at the interplation nodes x_0 and $x_1 = x_0 + h$:

$$\int\limits_{x_0}^{x_0+h} f(x)\,dx \approx \frac{h}{2}(y_0 + y_1) + \frac{h^2}{12}(y_0' - y_1'), \quad |R| \le \frac{h^5}{720} M_4. \tag{19.79}$$

The *Hermite trapezoidal formula* is obtained by summation:

$$\int\limits_a^b f(x)\,dx \approx h\left(\frac{y_0}{2} + y_1 + y_2 + \cdots + y_{n-1} + \frac{y_n}{2}\right) + \frac{h^2}{12}(y_0' - y_n'), \quad |R| \le \frac{(b-a)h^4}{720} M_4. \tag{19.80}$$

M_4 denotes an upper bound for $|f^{(4)}(x)|$ on the entire integration domain. The Hermite trapezoidal formula has an error of order 4 and it is exact for polynomials up to third degree.

19.3.3 Quadrature Formulas of Gauss

Quadrature formulas of Gauss have the general form

$$\int\limits_a^b f(x)\,dx \approx \sum_{\nu=0}^{n} c_\nu y_\nu \ \text{ with } \ y_\nu = f(x_\nu) \tag{19.81}$$

where not only the coefficients c_ν are considered as parameters but also the interpolation nodes x_ν. These parameters are determined in order to make the formula (19.81) exact for polynomials of the highest possible degree.
The quadrature formulas of Gauss result in very accurate approximations, but the interpolation nodes must be chosen in a very special way.

19.3.3.1 Gauss Quadrature Formulas

If the integration interval in (19.81) is chosen as $[a, b] = [-1, 1]$, and the interpolation nodes are chosen as the roots of the Legendre polynomials (see 9.1.2.6, **3.**, p. 566, 21.12, p. 1108), then the coefficients c_ν can be determined so that the formula (19.81) gives the exact value for polynomials up to degree $2n + 1$. The roots of the Legendre polynomials are symmetric with respect to the origin. For the cases $n = 1, 2$ and 3 they are:

$$\begin{array}{llll}
n=1: & x_0=-x_1, & c_0=1, & \\
 & x_1=\dfrac{1}{\sqrt{3}}=0.577\,350\,269\ldots, & c_1=1. & \\
n=2: & x_0=-x_2, & c_0=\dfrac{5}{9}, & \\
 & x_1=0, & c_1=\dfrac{8}{9}, & (19.82)\\
 & x_2=\sqrt{\dfrac{3}{5}}=0.774\,596\,669\ldots, & c_2=c_0. & \\
n=3: & x_0=-x_3, & c_0=0.347\,854\,854\ldots, & \\
 & x_1=-x_2, & c_1=0.652\,145\,154\ldots, & \\
 & x_2=0.339\,981\,043\ldots, & c_2=c_1, & \\
 & x_3=0.861\,136\,311\ldots, & c_3=c_0. &
\end{array}$$

Remark: A general integration interval $[a, b]$ can be transformed into $[-1, 1]$ by the transformation $t = \frac{b-a}{2}x + \frac{a+b}{2}$ $(t \in [a, b], x \in [-1, 1])$. Then

$$\int_a^b f(t)\,dt \approx \frac{b-a}{2}\sum_{\nu=0}^{n} c_\nu f\left(\frac{b-a}{2}x_\nu + \frac{a+b}{2}\right) \qquad (19.83)$$

with the values x_ν and c_ν given above for the interval $[-1, 1]$.

19.3.3.2 Lobatto's Quadrature Formulas

In some cases it is reasonable also to choose the endpoints of the subintervals as interpolation nodes. Then, there are $2n$ more free parameters in (19.81). These values can be determined so that polynomials up to degree $2n-1$ can be integrated exactly. In the cases $n = 2$ and $n = 3$:

$n = 2$:

$$x_0=-1,\ c_0=\frac{1}{3},\quad x_1=0,\ c_1=\frac{4}{3},\quad x_2=1,\ c_2=c_0. \qquad (19.84a)$$

$n = 3$:

$$\begin{array}{ll}
x_0=-1, & c_0=\dfrac{1}{6},\\
x_1=-x_2, & c_1=\dfrac{5}{6},\\
x_2=\dfrac{1}{\sqrt{5}}=0.447\,213\,595\ldots, & c_2=c_1,\\
x_3=1, & c_3=c_0.
\end{array} \qquad (19.84b)$$

The case $n = 2$ represents the Simpson formula.

19.3.4 Method of Romberg

To increase the accuracy of numerical integration the method of Romberg can be recommended, where one starts with a sequence of trapezoid sums, which is obtained by repeated halving of the integration step size.

19.3.4.1 Algorithm of the Romberg Method

The method consists of the following steps:

1. Trapezoid sums determination

The trapezoid sum $T(h_i)$ according to (19.76) in 19.3.2.2, p. 964 is determined as an approximation of the integral $\int_a^b f(x)\,dx$ with the step sizes

$$h_i = \frac{b-a}{2^i} \quad (i = 0, 1, 2, \ldots, m). \qquad (19.85)$$

Here, the recursive relation

$$T(h_i) = T\left(\frac{h_{i-1}}{2}\right) = \frac{h_{i-1}}{2}\left[\frac{1}{2}f(a) + f\left(a + \frac{h_{i-1}}{2}\right) + f(a+h_{i-1}) + f\left(a + \frac{3}{2}h_{i-1}\right)\right.$$
$$\left. + f(a+2h_{i-1}) + \cdots + f\left(a + \frac{2n-1}{2}h_{i-1}\right) + \frac{1}{2}f(b)\right] \tag{19.86}$$
$$= \frac{1}{2}T(h_{i-1}) + \frac{h_{i-1}}{2}\sum_{j=0}^{n-1} f\left(a + \frac{h_{i-1}}{2} + jh_{i-1}\right) \quad (i = 1, 2, \ldots, m;\ n = 2^{i-1})$$

is considered. Recursion formula (19.86) tells that for the calculation of $T(h_i)$ from $T(h_{i-1})$ the function values must be calculated only at the new interpolation nodes.

2. Triangular Scheme

$T_{0i} = T(h_i)\ \ (i = 0, 1, 2, \ldots)$ is substituted and the values

$$T_{ki} = T_{k-1,i} + \frac{T_{k-1,i} - T_{k-1,i-1}}{4^k - 1} \quad (k = 1, 2, \ldots, m;\ i = k, k+1, \ldots) \tag{19.87}$$

are calculated recursively. The arrangement of the values calculated according to (19.87) is most practical in a triangular scheme, whose elements are calculated in a column-wise manner:

$$\begin{array}{llll} T(h_0) = T_{00} & & & \\ T(h_1) = T_{01} & T_{11} & & \\ T(h_2) = T_{02} & T_{12} & T_{22} & \\ T(h_3) = T_{03} & T_{13} & T_{23} & T_{33} \\ \ldots\ldots\ldots & & & \end{array} \tag{19.88}$$

The scheme will be continued downwards (with a fixed number of columns) until the lower values at the right are almost the same. The values $T_{1i}\ \ (i = 1, 2, \ldots)$ of the second column correspond to those calculated by the Simpson formula.

19.3.4.2 Extrapolation Principle

The Romberg method represents an application of the so-called *extrapolation principle*. This will be demonstrated by deriving the formula (19.87) for the case $k = 1$. The required integral is denoted by I, the corresponding trapezoid sum (19.76) by $T(h)$. If the integrand of I is $(2m+2)$ times continuously differentiable in the integration interval, then it can be shown that an *asymptotical expansion* with respect to h is valid for the error R of the quadrature formula, and it has the form

$$R(h) = I - T(h) = a_1h^2 + a_2h^4 + \cdots + a_mh^{2m} + O(h^{2m+2}) \tag{19.89a}$$

or

$$T(h) = I - a_1h^2 - a_2h^4 - \cdots - a_mh^{2m} + O(h^{2m+2}). \tag{19.89b}$$

The coefficients $a_1, a_2, \ldots, a_m$ are constants and independent of h.

$T(h)$ and $T\left(\frac{h}{2}\right)$ are formed according to (19.89b) and the linear combination

$$T_1(h) = \alpha_1T(h) + \alpha_2T\left(\frac{h}{2}\right) = (\alpha_1+\alpha_2)I - a_1\left(\alpha_1 + \frac{\alpha_2}{4}\right)h^2 - a_2\left(\alpha_1 + \frac{\alpha_2}{16}\right)h^4 - \cdots \tag{19.90}$$

is considered. If $\alpha_1 + \alpha_2 = 1$ and $\alpha_1 + \frac{\alpha_2}{4} = 0$ are substituted, then $T_1(h)$ has an error of order 4, while $T(h)$ and $T(h/2)$ both have errors of order only 2. The formula is

$$T_1(h) = -\frac{1}{3}T(h) + \frac{4}{3}T\left(\frac{h}{2}\right) = T\left(\frac{h}{2}\right) + \frac{T\left(\frac{h}{2}\right) - T(h)}{3}. \tag{19.91}$$

This is the formula (19.87) for $k = 1$. Repeated application of the above procedure results in the approximation T_{ki} according to (19.87) and

$$T_{ki} = I + O(h_i^{2k+2}). \tag{19.92}$$

■ The definite integral $I = \int_0^1 \frac{\sin x}{x}\,dx$ (integral sine, see 8.2.5, **1.**, p. 513) cannot be obtained in an elementary way. Calculate the approximate values of this integral (calculating for 8 digits).

1. Romberg method:

$k = 0$	$k = 1$	$k = 2$	$k = 3$
0.92073549			
0.93979328	0.94614588		
0.94451352	0.94608693	0.94608300	
0.94569086	0.94608331	0.94608307	0.94608307.

The Romberg method results in the approximation value 0.94608307. The value calculated for 10 digits is 0.9460830704. The order $O\left((1/8)^8\right) \approx 6 \cdot 10^{-8}$ of the error according to (19.92) is verified.

2. Trapezoidal and Simpson Formulas: From the scheme of the Romberg method it can be got directly that for $h_3 = 1/8$ the trapezoid formula has the approximation value 0.94569086 and the Simpson formula gives the value 0.94608331.
The correction of the trapezoidal formula by Hermite according to (19.79) results in the value $I \approx 0.94569086 + \frac{0.30116868}{64 \cdot 12} = 0.94608301.$

3. Gauss Formula: By the formula (19.83) we get for

$$n = 1: \quad I \approx \frac{1}{2}\left[c_0 f\left(\frac{1}{2}x_0 + \frac{1}{2}\right) + c_1 f\left(\frac{1}{2}x_1 + \frac{1}{2}\right)\right] = 0.94604113;$$

$$n = 2: \quad I \approx \frac{1}{2}\left[c_0 f\left(\frac{1}{2}x_0 + \frac{1}{2}\right) + c_1 f\left(\frac{1}{2}x_1 + \frac{1}{2}\right) + c_2 f\left(\frac{1}{2}x_2 + \frac{1}{2}\right)\right] = 0.94608313;$$

$$n = 3: \quad I \approx \frac{1}{2}\left[c_0 f\left(\frac{1}{2}x_0 + \frac{1}{2}\right) + \cdots + c_3 f\left(\frac{1}{2}x_3 + \frac{1}{2}\right)\right] = 0.94608307.$$

It can be observed that the Gauss formula results in an 8-digit exact approximation value for $n = 3$, i.e., with only four function values. With the trapezoidal rule this accuracy would need a very large number (> 1000) of function values.

Remarks:

1. *Fourier analysis* has an important role in integrating periodic functions (see 7.4.1.1, **1.**, p. 474). The details of numerical realizations can be found under the title of *harmonic analysis* (see 19.6.4, p. 992). The actual computations are based on the so-called *Fast Fourier Transformation* FFT (see 19.6.4.2, p. 993).

2. In many applications it is useful to take the special properties of the integrands under consideration. Further integration routines can be developed for such special cases. A large variety of convergence properties, error analysis, and optimal integration formulas is discussed in the literature (see, e.g., [19.4]).

3. Numerical methods to find the values of *multiple integrals* are discussed in the literature (see, e.g., [19.26]).

19.4 Approximate Integration of Ordinary Differential Equations

In many cases, the solution of an ordinary differential equation cannot be given in closed form as an expression of known elementary functions. The solution, which still exists under rather general circumstances (see 9.1.1.1, p. 540), must be determined by numerical methods. These result only in particular solutions, but it is possible to reach high accuracy. Since differential equations of higher order than one can be either initial value problems or boundary value problems, numerical methods were developed for both types of problems.

19.4.1 Initial Value Problems

The principle of the methods presented in the following discussion to solve initial value problems

$$y' = f(x,y), \quad y(x_0) = y_0 \tag{19.93}$$

is to give approximate values y_i for the unknown function $y(x)$ at a chosen set of interpolation points x_i. Usually, *equidistant interpolation nodes* are considered with a previously given *step size* h:

$$x_i = x_0 + ih \quad (i = 0,1,2,\ldots). \tag{19.94}$$

19.4.1.1 Euler Polygonal Method

An integral representation of the initial value problem (19.93) is given by integration

$$y(x) = y_0 + \int_{x_0}^{x} f(x, y(x))\,dx. \tag{19.95}$$

This is the starting point for the approximation

$$y(x_1) = y_0 + \int_{x_0}^{x_0+h} f(x, y(x))\,dx \approx y_0 + hf(x_0, y_0) = y_1, \tag{19.96}$$

which is generalized as the *Euler broken line method* or *Euler polygonal method*:

$$y_{i+1} = y_i + hf(x_i, y_i) \quad (i = 0,1,2,\ldots;\ y(x_0) = y_0). \tag{19.97}$$

For a geometric interpretation see **Fig. 19.5**. Comparison of (19.96) with the Taylor expansion

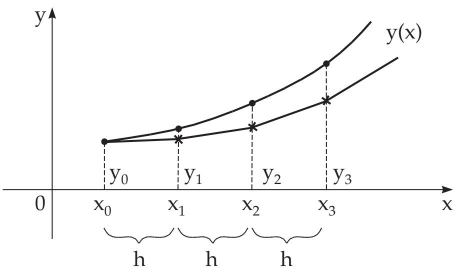


Figure 19.5

$$y(x_1) = y(x_0 + h) = y_0 + f(x_0, y_0)h + \frac{y''(\xi)}{2}h^2 \tag{19.98}$$

with $x_0 < \xi < x_0 + h$ shows that the approximation y_1 has an error of order h^2. The accuracy can be improved by reducing the step size h. Practical calculations show that halving the step size h results in halving the error of the approximations y_i.
A quick overview of the approximate shape of the solution curve can be got by using the Euler method.

19.4.1.2 Runge-Kutta Methods

1. Calculation Scheme

The equation $y'(x) = f(x,y)$ determines at every point (x_0, y_0) a direction, the direction of the tangent line of the solution curve passing through the point (x_0, y_0). The Euler method follows this direction until the next interpolation node. The Runge-Kutta methods consider more points "between" (x_0, y_0)

and the possible next point (x_0+h, y_1) of the curve, and depending on the appropriate choice of these additional points more accurate values are obtained for y_1. There exist Runge-Kutta methods of different orders depending on the number and the arrangements of these "auxiliary" points. Here a fourth-order method (see 19.4.1.5, **1.**, p. 972) is shown. (The Euler method is a first-order Runge-Kutta method.)

The calculation scheme of fourth order for the step from x_0 to $x_1 = x_0 + h$ to get an approximate value for y_1 of (19.93) is given in (19.99). The further steps follow the same scheme.

The error of this Runge-Kutta method has order h^5 (at every step) according to (19.99), so with an appropriate choice of the step size high accuracy can be obtained.

x	y	$k = h \cdot f(x,y)$
x_0	y_0	k_1
$x_0 + h/2$	$y_0 + k_1/2$	k_2
$x_0 + h/2$	$y_0 + k_2/2$	k_3
$x_0 + h$	$y_0 + k_3$	k_4
$x_1 = x_0 + h$	$y_1 = y_0 + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4)$	

(19.99)

■ $y' = \frac{1}{4}(x^2 + y^2)$ with $y(0) = 0$. $y(0.5)$ is determined in one step, i.e. $h = 0.5$ (see the table on the right). The exact value for 8 digits is 0.01041860.

x	y	$k = \frac{1}{8}(x^2 + y^2)$
0	0	0
0.25	0	0.00781250
0.25	0.00390625	0.00781441
0.5	0.00781441	0.03125763
0.5	0.01041858	

2. Remarks

1. For the special differential equation $y' = f(x)$, this Runge-Kutta method becomes the Simpson formula (see 19.3.2.3, p. 965).

2. For a large number of integration steps, a change of step size is possible or sometimes necessary. The change of step size can be decided by checking the accuracy so that the step is repeated with a double step size $2h$. If, e.g., the approximate value is $y_2(h)$ for $y(x_0+2h)$ (calculated by the single step size) and $y_2(2h)$ (calculated by the doubled step size), then the estimation for the error $R_2(h) = y(x_0+2h) - y_2(h)$ is

$$R_2(h) \approx \frac{1}{15}[y_2(h) - y_2(2h)]. \tag{19.100}$$

Information about the implementation of the step size changes can be found in the literature (see [19.24]).

3. Runge-Kutta methods can easily be used also for higher-order differential equations, see [19.24]. Higher-order differential equations can be rewritten in a first-order differential equation system (see p. 550). Then, the approximation methods are performed as parallel calculations according to (19.99), as the differential equations are connected to each other.

19.4.1.3 Multi-Step Methods

The Euler method (19.97) and the Runge-Kutta method (19.99) are so-called *single-step methods*, since we start only from y_i in the calculation of y_{i+1}. In general, *linear multi-step methods* have the form

$$\begin{aligned} &y_{i+k} + \alpha_{k-1}y_{i+k-1} + \alpha_{k-2}y_{i+k-2} + \cdots + \alpha_1 y_{i+1} + \alpha_0 y_i \\ &= h(\beta_k f_{i+k} + \beta_{k-1} f_{i+k-1} + \cdots + \beta_1 f_{i+1} + \beta_0 f_i) \end{aligned} \tag{19.101}$$

with appropriately chosen constants α_j and β_j ($j = 0, 1, \ldots, k$; $\alpha_k = 1$). The formula (19.101) is called a *k-step method* if $|\alpha_0| + |\beta_0| \neq 0$. It is called *explicit*, if $\beta_k = 0$, since in this case the values $f_{i+j} = f(x_{i+j}, y_{i+j})$ on the right-hand side of (19.101) only contain the already known approximation values $y_i, y_{i+1}, \ldots, y_{i+k-1}$. If $\beta_k \neq 0$ holds, the method is called *implicit*, since then the required new value y_{i+k} occurs on both sides of (19.101).

The k initial values $y_0, y_1, \ldots, y_{k-1}$ must be known in the application of a k-step method. These initial values can be got, e.g., by one-step methods.

A special multi-step method to solve the initial value problem (19.93) can be derived if the derivative

$y'(x_i)$ in (19.93) is replaced by a *difference formula* (see 9.1.1.5, **1.**, p. 549) or if the integral in (19.95) is approximated by a *quadrature formula* (see 19.3.1, p. 963).

Examples of special multi-step methods are:

1. Midpoint Rule The derivative $y'(x_{i+1})$ in (19.93) is replaced by the slope of the secant line between the interpolation nodes x_i and x_{i+2}, i.e.:

$$y_{i+2} - y_i = 2hf_{i+1}. \tag{19.102}$$

2. Rule of Milne The integral in (19.95) is approximated by the Simpson formula:

$$y_{i+2} - y_i = \frac{h}{3}(f_i + 4f_{i+1} + f_{i+2}). \tag{19.103}$$

3. Rule of Adams and Bashforth The integrand in (19.95) is replaced by the interpolation polynomial of Lagrange (see 19.6.1.2, p. 983) based on the k interpolation nodes $x_i, x_{i+1}, \ldots, x_{i+k-1}$. Integrating between x_{i+k-1} and x_{i+k} results in

$$y_{i+k} - y_{i+k-1} = \sum_{j=0}^{k-1}\left[\int\limits_{x_{i+k-1}}^{x_{i+k}} L_j(x)\,dx\right] f(x_{i+j}, y_{i+j}) = h\sum_{j=0}^{k-1}\beta_j f(x_{i+j}, y_{i+j}). \tag{19.104}$$

The method (19.104) is explicit for y_{i+k}. For the calculation of the coefficients β_j see [19.2].

19.4.1.4 Predictor-Corrector Method

In practice, implicit multi-step methods have a great advantage compared to explicit ones in that they allow much larger step sizes with the same accuracy. But, an implicit multi-step method usually requires the solution of a non-linear equation to get the approximation value y_{i+k}. This follows from (19.101) and has the form

$$y_{i+k} = h\sum_{j=0}^{k}\beta_j f_{i+j} - \sum_{j=0}^{k-1}\alpha_j y_{i+j} = F(y_{i+k}). \tag{19.105}$$

The solution of (19.105) is an iterative one. The procedure is the following: An initial value $y_{i+k}^{(0)}$ is determined by an explicit formula, the so-called *predictor*. Then it will be corrected by an iteration rule

$$y_{i+k}^{(\mu+1)} = F(y_{i+k}^{(\mu)}) \quad (\mu = 0, 1, 2, \ldots), \tag{19.106}$$

which is called the *corrector* coming from the implicit method. Special predictor–corrector formulas are:

1. $$y_{i+1}^{(0)} = y_i + \frac{h}{12}(5f_{i-2} - 16f_{i-1} + 23f_i), \tag{19.107a}$$

$$y_{i+1}^{(\mu+1)} = y_i + \frac{h}{12}(-f_{i-1} + 8f_i + 5f_{i+1}^{(\mu)}) \quad (\mu = 0, 1, \ldots); \tag{19.107b}$$

2. $$y_{i+1}^{(0)} = y_{i-2} + 9y_{i-1} - 9y_i + 6h(f_{i-1} + f_i), \tag{19.108a}$$

$$y_{i+1}^{(\mu+1)} = y_{i-1} + \frac{h}{3}(f_{i-1} + 4f_i + f_{i+1}^{(\mu)}) \quad (\mu = 0, 1, \ldots). \tag{19.108b}$$

The Simpson formula as the corrector in (19.108b) is numerically unstable and it can be replaced, e.g., by

$$y_{i+1}^{(\mu+1)} = 0.9y_{i-1} + 0.1y_i + \frac{h}{24}(0.1f_{i-2} + 6.7f_{i-1} + 30.7f_i + 8.1f_{i+1}^{(\mu)}). \tag{19.109}$$

19.4.1.5 Convergence, Consistency, Stability

1. Global Discretization Error and Convergence

Single-step methods can be written generally in the form:

$$y_{i+1} = y_i + hF(x_i, y_i, h) \quad (i = 0, 1, 2, \ldots;\ y_0 \text{ given}). \tag{19.110}$$

Here $F(x, y, h)$ is called the *increment function* or progressive direction of the single-step method. The approximating solution obtained by (19.110) depends on the step size h and it should be denoted by $y(x, h)$. Its difference from the exact solution $y(x)$ of the initial value problem (19.93) is called the *global discretization error* $g(x, h)$ (see (19.111)), and we say: The single-step method (19.110) is *convergent with order* p if p is the largest natural number such that

$$g(x, h) = y(x, h) - y(x) = O(h^p) \tag{19.111}$$

holds. Formula (19.111) says that the approximation $y(x, h)$ determined with the step size $h = \dfrac{x - x_0}{n}$ converges to the exact solution $y(x)$ for every x from the domain of the initial value problem if $h \to 0$.

■ The Euler method (19.97) has order of convergence $p = 1$. For the Runge-Kutta method (19.99) $p = 4$ holds.

2. Local Discretization Error and Consistency

The order of convergence according to (19.111) shows how well the approximating solution $y(x, h)$ approximates the exact solution $y(x)$. Beside this, it is an interesting question of how well the increment function $F(x, y, h)$ approximates the derivative $y' = f(x, y)$. For this purpose the so-called *local discretization error* $l(x, h)$ (see (19.112)) is introduced. The single-step method (19.110) is *consistent with order* p, if p is the largest natural number with

$$l(x, h) = \frac{y(x + h) - y(x)}{h} - F(x, y, h) = O(h^p). \tag{19.112}$$

It follows directly from (19.112) that for a consistent single-step method

$$\lim_{h \to 0} F(x, y, h) = f(x, y). \tag{19.113}$$

■ The Euler method has order of consistency $p = 1$, the Runge-Kutta method in (19.99) has order of consistency $p = 4$.

3. Stability with Respect to Perturbation of the Initial Values

In the practical performance of a single-step method, a rounding error $O(1/h)$ adds to the global discretization error $O(h^p)$. Consequently, we have to select a not too small, finite step size $h > 0$. It is also an important question of how the numerical solution y_i behaves under perturbations of the initial values or in the case $x_i \to \infty$.

In the theory of ordinary differential equations, an initial value problem (19.93) is called *stable with respect to perturbations of its initial values* if:

$$|\tilde{y}(x) - y(x)| \le |\tilde{y}_0 - y_0|. \tag{19.114}$$

Here $\tilde{y}(x)$ is the solution of (19.93) with the perturbed initial value $\tilde{y}(x_0) = \tilde{y}_0$ instead of y_0. Estimation (19.114) tells that the absolute value of the difference of the solutions is not larger than the perturbation of the initial values.

In general, it is hard to check (19.114). Therefore the *linear test problem*

$$y' = \lambda y \ \text{ with } \ y(x_0) = y_0 \quad (\lambda \ \text{constant}, \ \lambda \le 0) \tag{19.115}$$

is considered which is stable, and a single-step method is applied to this special initial value problem. A consistent method is called *absolutely stable* with step size $h > 0$ with respect to perturbed initial values if the approximating solution y_i of the above linear test problem (19.115) obtained by using the method satisfies the condition

$$|y_i| \le |y_0|. \tag{19.116}$$

■ Applying the Euler polygon method for equation (19.115) results in the solution $y_{i+1} = (1 + \lambda h)y_i$ $(i = 0, 1, \ldots)$. Obviously, (19.116) holds if $|1 + \lambda h| \le 1$, and so the step size must satisfy $-2 \le \lambda h \le 0$.

4. Stiff Differential Equations

Many application problems, including those in chemical kinetics, can be modeled by differential equations whose solutions consist of terms converging to zero exponentially but in a high different kind of exponential decreasing. These equations are called *stiff differential equations*. For example:

$$y(x) = C_1 e^{\lambda_1 x} + C_2 e^{\lambda_2 x} \quad (C_1, C_2, \lambda_1, \lambda_2 \text{ const}) \tag{19.117}$$

with $\lambda_1 < 0$, $\lambda_2 < 0$ and $|\lambda_1| \ll |\lambda_2|$, e.g., $\lambda_1 = -1$, $\lambda_2 = -1000$. The term with λ_2 does not have a significant affect on the solution function, but it does in selecting the step size h for a numerical method. In such cases the choice of the most appropriate numerical method has special importance (see [19.23]).

19.4.2 Boundary Value Problems

The most important methods for solving boundary value problems of ordinary differential equations will be demonstrated on the following simple linear boundary value problem for a differential equation of the second order:

$$y''(x) + p(x)y'(x) + q(x)y(x) = f(x) \quad (a \le x \le b) \text{ with } y(a) = \alpha,\ y(b) = \beta. \tag{19.118}$$

The functions $p(x)$, $q(x)$ and $f(x)$ and also the constants α and β are given.
The given method can also be adapted for boundary value problems of higher-order differential equations.

19.4.2.1 Difference Method

The interval $[a, b]$ is subdivided by equidistant interpolation points $x_\nu = x_0 + \nu h$ $(\nu = 0, 1, 2, \ldots, n;$ $x_0 = a,\ x_n = b)$ and the values of the derivatives are substituted into the differential equation at the interior interpolation points

$$y''(x_\nu) + p(x_\nu)y'(x_\nu) + q(x_\nu)y(x_\nu) = f(x_\nu) \quad (\nu = 1, 2, \ldots, n-1) \tag{19.119}$$

by so-called *finite divided differences*, e.g.:

$$y'(x_\nu) \approx y'_\nu = \frac{y_{\nu+1} - y_{\nu-1}}{2h}, \tag{19.120a}$$

$$y''(x_\nu) \approx y''_\nu = \frac{y_{\nu+1} - 2y_\nu + y_{\nu-1}}{h^2}. \tag{19.120b}$$

In this way $n-1$ linear equations are obtained for the $n-1$ approximation values $y_\nu \approx y(x_\nu)$ in the interior of the integration interval $[a, b]$, considering the conditions $y_0 = \alpha$ and $y_n = \beta$. If the boundary conditions also contain derivatives, they must also be replaced by finite expressions.

Eigenvalue problems of differential equations (see 9.1.3.2, p. 569) are handled analogously. The application of the *difference method*, described by (19.119) and (19.120a,b), leads to a matrix eigenvalue problem (see 4.6, p. 314).

■ The solution of the homogeneous differential equation $y'' + \lambda^2 y = 0$ with boundary conditions $y(0) = y(1) = 0$ leads to a matrix eigenvalue problem. The difference method transforms the differential equation into the difference equation $y_{\nu+1} - 2y_\nu + y_{\nu-1} + h^2\lambda^2 y_\nu = 0$. If three interior points are chosen, hence $h = 1/4$, then considering $y_0 = y(0) = 0$, $y_4 = y(1) = 0$ the discrete system is

$$\begin{aligned}
\left(-2 + \frac{\lambda^2}{16}\right) y_1 + y_2 \qquad\qquad &= 0,\\
y_1 + \left(-2 + \frac{\lambda^2}{16}\right) y_2 + y_3 &= 0,\\
y_2 + \left(-2 + \frac{\lambda^2}{16}\right) y_3 &= 0.
\end{aligned}$$

This homogeneous system of equations has a non-trivial solution only when the coefficient determinant

is zero. This condition results in the eigenvalues ${\lambda_1}^2 = 9.37$, ${\lambda_2}^2 = 32$ and ${\lambda_3}^2 = 54.63$. Among them only the smallest one is close to its corresponding true value 9.87.

Remark: The accuracy of the difference method can be improved by

1. decreasing the step size h,

2. application of a derivative approximation of higher order (approximations as (19.120a,b) have an error of order $O(h^2)$),

3. application of multi-step methods (see 19.4.1.3, p. 970).

If the problem is a non-linear boundary value problem, then the difference method leads to a system of non-linear equations of the unknown approximation values y_ν (see 19.2.2, p. 961).

19.4.2.2 Approximation by Using Given Functions

The approximate solution of the boundary value problem (19.118) is a linear combination of suitably chosen functions $g_i(x)$, which are linearly independent and each one satisfies the boundary value conditions:

$$y(x) \approx g(x) = \sum_{i=1}^{n} a_i g_i(x). \qquad (19.121)$$

Substituting $g(x)$ into the differential equation (19.118) results in an error, the so-called *defect*

$$\varepsilon(x; a_1, a_2, \ldots, a_n) = g''(x) + p(x)g'(x) + q(x)g(x) - f(x). \qquad (19.122)$$

To determine the coefficients a_i the following principles (see also p. 978) can be used:

1. Collocation Method The defect has to be zero at n given points x_ν, the so-called *collocation points*. The conditions

$$\varepsilon(x_\nu; a_1, a_2, \ldots, a_n) = 0 \quad (\nu = 1, 2, \ldots, n), \quad a < x_1 < x_2 < \ldots < x_n < b \qquad (19.123)$$

result in a linear system of equations for the unknown coefficients.

2. Least Squares Method The integral

$$F(a_1, a_2, \ldots, a_n) = \int_a^b \varepsilon^2(x; a_1, a_2, \ldots, a_n)\, dx, \qquad (19.124)$$

depending on the coefficients, should be minimal. The necessary conditions

$$\frac{\partial F}{\partial a_i} = 0 \quad (i = 1, 2, \ldots, n) \qquad (19.125)$$

give a linear system of equations for the coefficients a_i.

3. Galerkin Method The requirement is that the so-called *error orthogonality* is satisfied, i.e.,

$$\int_a^b \varepsilon(x; a_1, a_2, \ldots, a_n) g_i(x)\, dx = 0 \quad (i = 1, 2, \ldots, n), \qquad (19.126)$$

and in this way a linear system of equations is obtained for the unknown coefficients.

4. Ritz Method The solution $y(x)$ often has the property that it minimizes the *variational integral*,

$$I[y] = \int_a^b H(x, y, y')\, dx \qquad (19.127)$$

(see (10.4), p. 610). If the function $H(x, y, y')$ is known, then $y(x)$ is replaced by the approximation $g(x)$ as in (19.121) and $I[y] = I(a_1, a_2, \ldots, a_n)$ is minimized. The necessary conditions

$$\frac{\partial I}{\partial a_i} = 0 \quad (i = 1, 2, \ldots, n) \qquad (19.128)$$

result in n equation for the coefficients a_i.

■ Under certain conditions on the functions p, q, f and y, the boundary value problem

$$-[p(x)y'(x)]' + q(x)y(x) = f(x) \text{ with } y(a) = \alpha,\ y(b) = \beta \tag{19.129}$$

and the variational problem

$$I[y] = \int_a^b [p(x)y'^2(x) + q(x)y^2(x) - 2f(x)y(x)]\, dx = \min! \text{ with } y(a) = \alpha,\ y(b) = \beta \tag{19.130}$$

are equivalent, so $H(x, y, y')$ can be got immediately from (19.130) for the boundary value problem of the form (19.129).
Instead of the approximation (19.121), one often considers

$$g(x) = g_0(x) + \sum_{i=1}^{n} a_i g_i(x), \tag{19.131}$$

where $g_0(x)$ satisfies the boundary values and the functions $g_i(x)$ satisfy the conditions

$$g_i(a) = g_i(b) = 0 \quad (i = 1, 2, \ldots, n). \tag{19.132}$$

For the problem (19.118), an appropriate choice is, e.g.,

$$g_0(x) = \alpha + \frac{\beta - \alpha}{b - a}(x - a). \tag{19.133}$$

Remark: In a linear boundary value problem, the forms (19.121) and (19.131) result in a linear system of equations for the coefficients. In the case of non-linear boundary value problems non-linear systems of equations are obtained, which can be solved by the methods given in Section 19.2.2, p. 961.

19.4.2.3 Shooting Method

With the shooting method, the solution of a boundary value problem is reduced to the solution of an initial value problem. The basic idea of the method is described below as the *single-target method*.

1. Single-Target Method

The initial value problem

$$y'' + p(x)y' + q(x)y = f(x) \text{ with } y(a) = \alpha,\ y'(a) = s \tag{19.134}$$

is associated to the boundary value problem (19.118). Here s is a parameter, from which the solution y of the initial-value problem (19.134) depends on, i.e., $y = y(x, s)$ holds. The function $y(x, s)$ satisfies the first boundary condition $y(a, s) = \alpha$ according to (19.134). The parameter s should be determined so that $y(x, s)$ satisfies the second boundary condition $y(b, s) = \beta$. Therefore, one has to solve the equation

$$F(s) = y(b, s) - \beta, \tag{19.135}$$

and the regula falsi (or secant) method is an appropriate method to do this. It needs only the values of the function $F(s)$, but the computation of every function value requires the solution of an initial value problem (19.134) until $x = b$ for the special parameter value s with one of the methods given in 19.4.1.

2. Multiple-Target Method

In a so-called *multiple-target method*, the integration interval $[a, b]$ is divided into subintervals, and we use the single-target method on every subinterval. Then, the required solution is composed from the solutions of the subintervals, where the continuous transition at the endpoints of the subintervals must be ensured.
This requirement results in further conditions. For the numerical implementation of the multiple-target method, which is used mostly for non-linear boundary value problems, see. [19.24].

19.5 Approximate Integration of Partial Differential Equations

In this section only the principles of numerical solutions of partial differential equations are discussed using the example of linear second-order partial differential equations with two independent variables with the corresponding boundary or/and initial conditions.

19.5.1 Difference Method

A regular grid is considered on the integration domain by the chosen points (x_μ, y_ν). Usually, this grid is chosen to be rectangular and equally spaced:

$$x_\mu = x_0 + \mu h, \quad y_\nu = y_0 + \nu l \quad (\mu, \nu = 1, 2, \ldots). \tag{19.136}$$

It results in squares for $l = h$. If the required solution is denoted by $u(x, y)$, then the partial derivatives occurring in the differential equation and in the boundary or initial conditions are replaced by *finite divided differences* in the following way, where $u_{\mu\nu}$ denotes an approximate value for the function value $u(x_\mu, y_\nu)$:

Partial Derivative	Finite Divided Difference	Order of Error
$\dfrac{\partial u}{\partial x}(x_\mu, y_\nu)$	$\dfrac{1}{h}(u_{\mu+1,\nu} - u_{\mu,\nu})$ or $\dfrac{1}{2h}(u_{\mu+1,\nu} - u_{\mu-1,\nu})$	$O(h)$ or $O(h^2)$
$\dfrac{\partial u}{\partial y}(x_\mu, y_\nu)$	$\dfrac{1}{l}(u_{\mu,\nu+1} - u_{\mu,\nu})$ or $\dfrac{1}{2l}(u_{\mu,\nu+1} - u_{\mu,\nu-1})$	$O(l)$ or $O(l^2)$
$\dfrac{\partial^2 u}{\partial x \partial y}(x_\mu, y_\nu)$	$\dfrac{1}{4hl}(u_{\mu+1,\nu+1} - u_{\mu+1,\nu-1} - u_{\mu-1,\nu+1} + u_{\mu-1,\nu-1})$	$O(hl)$
$\dfrac{\partial^2 u}{\partial x^2}(x_\mu, y_\nu)$	$\dfrac{1}{h^2}(u_{\mu+1,\nu} - 2u_{\mu,\nu} + u_{\mu-1,\nu})$	$O(h^2)$
$\dfrac{\partial^2 u}{\partial y^2}(x_\mu, y_\nu)$	$\dfrac{1}{l^2}(u_{\mu,\nu+1} - 2u_{\mu,\nu} + u_{\mu,\nu-1})$	$O(l^2)$

(19.137)

The error order in (19.137) is given by using the Landau symbol O.
In some cases, it is more practical to apply the approximation

$$\frac{\partial^2 u}{\partial x^2}(x_\mu, y_\nu) \approx \sigma \frac{u_{\mu+1,\nu+1} - 2u_{\mu,\nu+1} + u_{\mu-1,\nu+1}}{h^2} + (1-\sigma)\frac{u_{\mu+1,\nu} - 2u_{\mu,\nu} + u_{\mu-1,\nu}}{h^2} \tag{19.138}$$

with a fixed parameter σ $(0 \le \sigma \le 1)$. Formula (19.138) represents a convex linear combination of two finite expressions obtained from the corresponding formula (19.137) for the values $y = y_\nu$ and $y = y_{\nu+1}$.

A partial differential equation can be rewritten as a *difference equation* at every interior point of the grid by the formulas (19.137), where the boundary and initial conditions are considered, as well. This system of equations for the approximation values $u_{\mu,\nu}$ has a large dimension for small step sizes h and l, so usually, it is solved by an iteration method (see 19.2.1.4, p. 960).

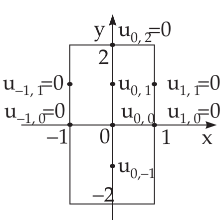


Figure 19.6

■ **A:** The function $u(x, y)$ should be the solution of the differential equation $\Delta u = u_{xx} + u_{yy} = -1$ for the points (x, y) with $|x| < 1$, $|y| < 2$, i.e., in the interior of a rectangle, and it should satisfy the boundary conditions $u = 0$ for $|x| = 1$ and $|y| = 2$. The difference equation corresponding to the differential equation for a square grid with step size h is: $4u_{\mu,\nu} = u_{\mu+1,\nu} + u_{\mu,\nu+1} + u_{\mu-1,\nu} + u_{\mu,\nu-1} + h^2$. The step size $h = 1$ (**Fig. 19.6**) results in a first rough approximation for the function values at the three interior points: $4u_{0,1} = 0 + 0 + 0 + u_{0,0} + 1$, $4u_{0,0} = 0 + u_{0,1} + 0 + u_{0,-1} + 1$, $4u_{0,-1} = 0 + u_{0,0} + 0 + 0 + 1$.

The solution is $u_{0,0} = \dfrac{3}{7} \approx 0.429$, $u_{0,1} = u_{0,-1} = \dfrac{5}{14} \approx 0.357$.

■ **B:** The system of equations arising in the application of the difference method for partial differential equations has a very special structure. It is demonstrated by the following example which is a more general boundary value problem. The integration domain is the square G: $0 \le x \le 1$, $0 \le y \le 1$. A function $u(x,y)$ should be determined for which $\Delta u = u_{xx} + u_{yy} = f(x,y)$ in the interior of G, $u(x,y) = g(x,y)$ on the boundary of G. The functions f and g are given. The difference equation associated to this differential equation is, for $h = l = 1/n$:

$$u_{\mu+1,\nu} + u_{\mu,\nu+1} + u_{\mu-1,\nu} + u_{\mu,\nu-1} - 4u_{\mu,\nu} = \frac{1}{n^2} f(x_\mu, y_\nu) \quad (\mu, \nu = 1, 2, \ldots, n-1).$$

In the case of $n = 5$, the left-hand side of this system of difference equations for the approximation values $u_{\mu,\nu}$ in the 4×4 interior points has the form (19.139)

$$\left(\begin{array}{cccc}
\boxed{\begin{matrix}-4&1&0&0\\1&-4&1&0\\0&1&-4&1\\0&0&1&-4\end{matrix}} & \boxed{\begin{matrix}1&0&0&0\\0&1&0&0\\0&0&1&0\\0&0&0&1\end{matrix}} & & \mathbf{O} \\
\boxed{\begin{matrix}1&0&0&0\\0&1&0&0\\0&0&1&0\\0&0&0&1\end{matrix}} & \boxed{\begin{matrix}-4&1&0&0\\1&-4&1&0\\0&1&-4&1\\0&0&1&-4\end{matrix}} & \boxed{\begin{matrix}1&0&0&0\\0&1&0&0\\0&0&1&0\\0&0&0&1\end{matrix}} & \\
 & \boxed{\begin{matrix}1&0&0&0\\0&1&0&0\\0&0&1&0\\0&0&0&1\end{matrix}} & \boxed{\begin{matrix}-4&1&0&0\\1&-4&1&0\\0&1&-4&1\\0&0&1&-4\end{matrix}} & \boxed{\begin{matrix}1&0&0&0\\0&1&0&0\\0&0&1&0\\0&0&0&1\end{matrix}} \\
\mathbf{O} & & \boxed{\begin{matrix}1&0&0&0\\0&1&0&0\\0&0&1&0\\0&0&0&1\end{matrix}} & \boxed{\begin{matrix}-4&1&0&0\\1&-4&1&0\\0&1&-4&1\\0&0&1&-4\end{matrix}}
\end{array}\right)
\left(\begin{array}{c} u_{11}\\u_{21}\\u_{31}\\u_{41}\\\hline u_{12}\\u_{22}\\u_{32}\\u_{42}\\\hline u_{13}\\u_{23}\\u_{33}\\u_{43}\\\hline u_{14}\\u_{24}\\u_{34}\\u_{44}\end{array}\right) \tag{19.139}$$

if the grid is considered row-wise from left to right, and considering that the values of the function are given on the boundary. The coefficient matrix is symmetric and is a *sparse matrix*. This form is called *block-tridiagonal*. It is obvious that the form of the matrix depends on how the grid-points are selected. For different classes of partial differential equations of second order, such as elliptic, parabolic and hyperbolic differential equations, more effective methods have been developed, and also the convergence and stability conditions have been investigated. There is a huge number of books about this topic (see, e.g., [19.22], [19.24]).

19.5.2 Approximation by Given Functions

The solution $u(x,y)$ is approximated by a function in the form

$$u(x,y) \approx v(x,y) = v_0(x,y) + \sum_{i=1}^{n} a_i v_i(x,y). \tag{19.140}$$

Here, two cases are distinguished:

1. $v_0(x,y)$ satisfies the given inhomogeneous differential equation, and the further functions $v_i(x,y)$ $(i = 1, 2, \ldots, n)$ satisfy the corresponding homogeneous differential equation (then the linear combination has to be found which approximates the given boundary conditions as well as possible).

2. $v_0(x,y)$ satisfies the inhomogeneous boundary conditions and the other functions $v_i(x,y)$ $(i = 1, 2, \ldots, n)$ satisfy the homogeneous boundary conditions (then the linear combination has to be found which approximates the solution of the differential equation on the considered domain as well as possible).

In both cases substituting the approximating function $v(x, y)$ from (19.140) in the first case into the boundary conditions, in the second case into the differential equation results in an error term, the so-called *defect*:

$$\varepsilon = \varepsilon(x, y; a_1, a_2, \ldots, a_n). \tag{19.141}$$

To determine the unknown coefficients a_i one of the following methods can be applied.

1. Collocation Method

The defect ε should be zero in n reasonably distributed points, at the *collocation points* (x_ν, y_ν) $(\nu = 1, 2, \ldots, n)$:

$$\varepsilon(x_\nu, y_\nu; a_1, a_2, \ldots, a_n) = 0 \quad (\nu = 1, 2, \ldots, n). \tag{19.142}$$

The collocation points in the first case are boundary points (it is called *boundary collocation*), in the second case they are interior points of the integration domain (it is called *domain collocation*).
From (19.142) are obtained n equations for the coefficients. Boundary collocation is usually preferred to domain collocation.

■ This method is applied to the example solved in 19.5.1 by the difference method, with the functions satisfying the differential equation:
$v(x, y; a_1, a_2, a_3) = -\frac{1}{4}(x^2 + y^2) + a_1 + a_2(x^2 - y^2) + a_3(x^4 - 6x^2y^2 + y^4)$.
The coefficients are determined to satisfy the boundary conditions at the points $(x_1, y_1) = (1, 0.5)$, $(x_2, y_2) = (1, 1.5)$ and $(x_3, y_3) = (0.5, 2)$ (boundary collocation). The linear system of equations

$$\begin{aligned} -0.3125 + a_1 + 0.75a_2 - 0.4375a_3 &= 0,\\ -0.8125 + a_1 - 1.25a_2 - 7.4375a_3 &= 0,\\ -1.0625 + a_1 - 3.75a_2 + 10.0625a_3 &= 0 \end{aligned}$$

has the solution $a_1 = 0.4562$, $a_2 = -0.2000$, $a_3 = -0.0143$. The approximate values of the solution can be calculated at arbitrary points with the approximating function. To compare the values with those obtained by the difference method: $v(0, 1) = 0.3919$ and $v(0, 0) = 0.4562$.

2. Least Squares Method

Depending on whether the approximation function (19.140) satisfies the differential equation or the boundary conditions, it is required

1. either the line integral over the boundary C

$$I = \int_{(C)} \varepsilon^2(x(t), y(t)\,;\, a_1, \ldots, a_n)\, dt = \min, \tag{19.143a}$$

where the boundary curve C is given by a parametric representation $x = x(t)$, $y = y(t)$,

2. or the double integral over the domain G

$$I = \iint_{(G)} \varepsilon^2(x, y;\, a_1, \ldots, a_n)\, dx\, dy = \min. \tag{19.143b}$$

From the necessary conditions, $\dfrac{\partial I}{\partial a_i} = 0$ $(i = 1, 2, \ldots, n)$, n equations are obtained for computing the parameters $a_1, a_2, \ldots, a_n$.

19.5.3 Finite Element Method (FEM)

After the appearance of modern computers the finite element methods became the most important technique for solving partial differential equations. These powerful methods give results which are easy to interpret.
Depending on the types of various applications, the FEM is implemented in very different ways, so here only the basic idea is given. It is similar to those used in the Ritz method (see 19.4.2.2, p. 974) for numerical solution of boundary value problems for ordinary differential equations and is related to spline approximations (see 19.7, p. 996).
The finite element method has the following steps:

1. Defining a Variational Problem A variational problem is formulated to the given boundary value problem. The process is demonstrated on the following boundary value problem:

$$\Delta u = u_{xx} + u_{yy} = f \text{ in the interior of } G,\ u = 0 \text{ on the boundary of } G. \tag{19.144}$$

The differential equation in (19.144) is multiplied by an appropriate smooth function $v(x, y)$ vanishing on the boundary of G, and it is integrated over the entire G to get

$$\iint\limits_{(G)} \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) v\, dx\, dy = \iint\limits_{(G)} f v\, dx\, dy. \tag{19.145}$$

Applying the Gauss integral formula (see 13.3.3.1, **2.**, p. 725), where $P(x, y) = -vu_y$ and $Q(x, y) = vu_x$ are substituted in (13.121), the *variational equation* from (19.145)

$$a(u, v) = b(v) \tag{19.146a}$$

is obtained with

$$a(u, v) = -\iint\limits_{(G)} \left(\frac{\partial u}{\partial x}\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\frac{\partial v}{\partial y}\right) dx\, dy, \quad b(v) = \iint\limits_{(G)} f v\, dx\, dy. \tag{19.146b}$$

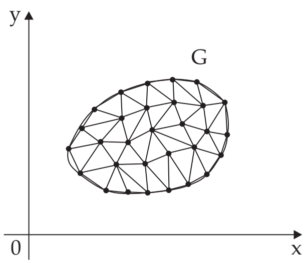


Figure 19.7

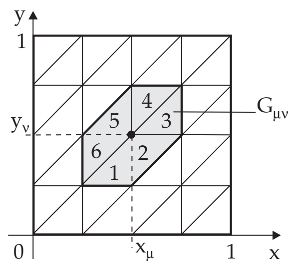


Figure 19.8

2. Triangularization The domain of integration G is decomposed into simple subdomains. Usually, a *triangularization* is used, where G is covered by triangles so that the neighboring triangles have a complete side or only a single vertex in common. Every domain bounded by curves can be approximated quite well by a union of triangles **(Fig. 19.7)**.

Remark: To avoid numerical difficulties, the triangularization should not contain obtuse-angled triangles.

■ A triangularization of the unit square could be performed as shown in **Fig. 19.8**. Here one starts from the grid points with coordinates $x_\mu = \mu h$, $y_\nu = \nu h$ $(\mu, \nu = 0, 1, 2, \ldots, N;\ h = 1/N)$. There are $(N-1)^2$ interior points. Considering the choice of the solution functions, it is always useful to consider the surface elements $G_{\mu\nu}$ composed of the six triangles having the common point (x_μ, y_ν). (In other cases, the number of triangles may differ from six. These surface elements are obviously not mutually exclusive.)

3. Solution A supposed approximating solution is defined for the required function $u(x, y)$ in every triangle. A triangle with the corresponding supposed solution is called a *finite element*. Polynomials in x and y are the most suitable choices. In many cases, the linear approximation

$$\tilde{u}(x, y) = a_1 + a_2 x + a_3 y \tag{19.147}$$

is sufficient. The supposed approximating function must be continuous under the transition from one triangle to neighboring ones, so a continuous final solution arises.

The coefficients a_1, a_2 and a_3 in (19.147) are uniquely defined by the values of the functions u_1, u_2 and u_3 at the three vertices of the triangle. The continuous transition to the neighboring triangles is

ensured by this at the same time. The supposed solution (19.147) contains the approximating values u_i of the required function as unknown parameters. For the supposed solution, which is applied as an approximation in the entire domain G for the required solution $u(x, y)$,

$$\tilde{u}(x, y) = \sum_{\mu=1}^{N-1} \sum_{\nu=1}^{N-1} \alpha_{\mu\nu} u_{\mu\nu}(x, y). \tag{19.148}$$

is chosen. The appropriate coefficients $\alpha_{\mu\nu}$ are determined. The following must be valid for the functions $u_{\mu\nu}(x, y)$: They represent a linear function over every triangle of $G_{\mu\nu}$ according to (19.147) with the following conditions:

1. $u_{\mu\nu}(x_k, y_l) = \begin{cases} 1 & \text{for } k = \mu, l = \nu, \\ 0 & \text{at any other grid point of } G_{\mu\nu}. \end{cases}$ (19.149a)

2. $u_{\mu\nu}(x, y) \equiv 0 \quad \text{for } (x, y) \notin G_{\mu\nu}.$ (19.149b)

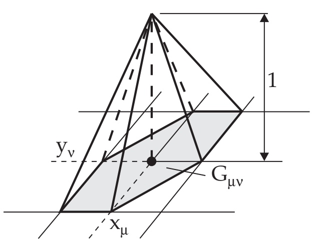


Figure 19.9

The representation of $u_{\mu\nu}(x, y)$ over $G_{\mu\nu}$ is shown in **Fig. 19.9**.
The calculation of $u_{\mu\nu}$ over $G_{\mu\nu}$, i.e., over all triangles 1 to 6 in **Fig. 19.8** is shown here only for triangle 1:

$$u_{\mu\nu}(x, y) = a_1 + a_2 x + a_3 \text{ with} \tag{19.150}$$

$$u_{\mu\nu}(x, y) = \begin{cases} 1 \text{ for } x = x_\mu, y = y_\nu\,, \\ 0 \text{ for } x = x_{\mu-1}, y = y_{\nu-1}, \\ 0 \text{ for } x = x_\mu, y = y_{\nu-1}. \end{cases} \tag{19.151}$$

From (19.151) $a_1 = 1 - \nu, a_2 = 0,\ a_3 = 1/h$, follow and so for triangle 1

$$u_{\mu\nu}(x, y) = 1 + \left(\frac{y}{h} - \nu\right). \tag{19.152}$$

Analogously, there is:

$$u_{\mu\nu}(x, y) = \begin{cases} 1 - \left(\frac{x}{h} - \mu\right) + \left(\frac{y}{h} - \nu\right) & \text{for triangle 2,} \\ 1 - \left(\frac{x}{h} - \mu\right) & \text{for triangle 3,} \\ 1 - \left(\frac{y}{h} - \nu\right) & \text{for triangle 4,} \\ 1 + \left(\frac{x}{h} - \mu\right) + \left(\frac{y}{h} - \nu\right) & \text{for triangle 5,} \\ 1 + \left(\frac{x}{h} - \mu\right) & \text{for triangle 6.} \end{cases} \tag{19.153}$$

4. Calculation of the Solution Coefficients The solution coefficients $\alpha_{\mu\nu}$ are determined by the requirements that the solution (19.148) satisfies the variational problem (19.146a) for every solution function $u_{\mu\nu}$, i.e., $\tilde{u}(x, y)$ is substituted for $u(x, y)$ and $u_{\mu\nu}(x, y)$ for $v(x, y)$ in (19.146a). In this way, a linear system of equations

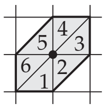
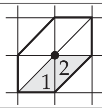
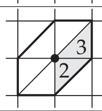
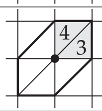
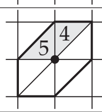
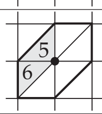
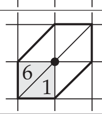
$$\sum_{\mu=1}^{N-1} \sum_{\nu=1}^{N-1} \alpha_{\mu\nu} a(u_{\mu\nu}, u_{kl}) = b(u_{kl}) \quad (k, l = 1, 2, \ldots, N-1) \tag{19.154}$$

is obtained for the unknown coefficients, where

$$a(u_{\mu\nu}, u_{kl}) = \iint\limits_{G_{kl}} \left(\frac{\partial u_{\mu\nu}}{\partial x}\frac{\partial u_{kl}}{\partial x} + \frac{\partial u_{\mu\nu}}{\partial y}\frac{\partial u_{kl}}{\partial y}\right) dx\, dy, \quad b(u_{kl}) = \iint\limits_{G_{kl}} f u_{kl}\, dx\, dy. \tag{19.155}$$

In the calculation of $a(u_{\mu\nu}, u_{kl})$ the fact must be kept in mind that integration is needed only in the cases of domains $G_{\mu\nu}$ and G_{kl} with non-empty intersection. These domains are denoted by shadowing in **Table 19.1**.

Table 19.1 Auxiliary table for FEM

Surface region	**Graphical representation**	**Triangle of G_{kl} $G_{\mu\nu}$**	$\frac{\partial u_{kl}}{\partial x}$	$\frac{\partial u_{\mu\nu}}{\partial x}$	$\sum \frac{\partial u_{kl}}{\partial x}\frac{\partial u_{\mu\nu}}{\partial x}$
1. $\mu = k$, $\nu = l$		1 1 2 2 3 3 4 4 5 5 6 6	0 $-1/h$ $-1/h$ 0 $1/h$ $1/h$	0 $-1/h$ $-1/h$ 0 $1/h$ $1/h$	$\frac{4}{h^2}$
2. $\mu = k$, $\nu = l-1$		1 5 2 4	0 $-1/h$	$1/h$ 0	0
3. $\mu = k+1$, $\nu = l$		2 6 3 5	$-1/h$ $-1/h$	$1/h$ $1/h$	$-\frac{2}{h^2}$
4. $\mu = k+1$, $\nu = l+1$		3 1 4 6	$-1/h$ 0	0 $1/h$	0
5. $\mu = k$, $\nu = l+1$		4 2 5 1	0 $-1/h$	$1/h$ 0	0
6. $\mu = k-1$, $\nu = l$		5 3 6 2	$1/h$ $1/h$	$-1/h$ $-1/h$	$-\frac{2}{h^2}$
7. $\mu = k-1$, $\nu = l-1$		6 4 1 3	$1/h$ 0	0 $-1/h$	0

The integration is always performed over a triangle with an area $h^2/2$, so for the partial derivatives with respect to x:

$$\frac{1}{h^2}\left(4\alpha_{kl} - 2\alpha_{k+1,l} - 2\alpha_{k-1,l}\right)\frac{h^2}{2} \tag{19.156a}$$

is obtained. Analogously, for the partial derivatives with respect to y the corresponding term is

$$\frac{1}{h^2}\left(4\alpha_{kl} - 2\alpha_{k,l+1} - 2\alpha_{k,l-1}\right)\frac{h^2}{2}. \tag{19.156b}$$

The calculation of the right-hand side $b(u_{kl})$ of (19.154) gives:

$$b(u_{kl}) = \iint\limits_{G_{kl}} f(x,y)u_{kl}(x,y)\,dx\,dy \approx f_{kl}V_P, \tag{19.157a}$$

where V_P is the volume of the pyramid over G_{kl} with height 1, determined by $u_{kl}(x,y)$ **(Fig. 19.9)**. Since

$$V_P = \frac{1}{3}\cdot 6\cdot\frac{1}{2}h^2, \quad \text{the approximation is} \quad b(u_{kl}) \approx f_{kl}h^2. \tag{19.157b}$$

So, the variational equations (19.154) result in the linear system of equations

$$4\alpha_{kl} - \alpha_{k+1,l} - \alpha_{k-1,l} - \alpha_{k,l+1} - \alpha_{k,l-1} = h^2 f_{kl} \quad (k,l = 1,2,\ldots,N-1) \tag{19.158}$$

for the determination of the solution coefficients.

Remarks:

1. If the solution coefficients are determined by (19.158), then $\tilde{u}(x,y)$ from (19.148) represents an explicit approximating solution, whose values can be calculated for an arbitrary point (x,y) from G.

2. If the integration domain must be covered by an irregular triangular grid then it is useful to introduce *triangular coordinates* (also called *barycentric coordinates*). In this way, the position of a point can be easily determined with respect to the triangular grid, and the calculation of the multidimensional integral is made easier as in (19.155), because every triangle can be easily transformed into the unit triangle with vertices $(0,0)$, $(0,1)$, $(1,0)$.

3. If accuracy must be improved or also the differentiability of the solution is required, then piecewise quadratic or cubic functions must be applied to obtain the supposed approximation (see, e.g., [19.22]).

4. In practical applications, usually systems of huge dimensions are obtained. This is the reason why so many special methods have been developed, e.g., for automatic triangularization and for practical enumeration of the elements (the structure of the system of equations depends on it). For detailed discussion of FEM see [19.13], [19.7], [19.22].

19.6 Approximation, Computation of Adjustment, Harmonic Analysis

19.6.1 Polynomial Interpolation

The basic problem of interpolation is to fit a curve through a sequence of points (x_ν, y_ν) $(\nu = 0,1,\ldots,n)$. This can happen graphically by any curve-fitting gadget, or numerically by a function $g(x)$, which takes given values y_ν at the points x_ν, at the so-called *interpolation points*. That is $g(x)$ satisfies the *interpolation conditions*

$$g(x_\nu) = y_\nu \quad (\nu = 0,1,2,\ldots,n). \tag{19.159}$$

In the first place, polynomials are used as interpolation functions, or for periodic functions so-called trigonometric polynomials. In this last case one talks about *trigonometric interpolation* (see 19.6.4.1, **2.**, p. 992). There are $n+1$ interpolation points, the order of the interpolation is n, and the highest degree of the interpolation polynomial is at most n. Since with increasing degree of the polynomials, strong oscillation may occur, which is usually not required, the interpolation interval can be decomposed into subintervals and a *spline interpolation* (see 19.7, p. 996) can be performed.

19.6.1.1 Newton's Interpolation Formula

To solve the interpolation problem (19.159) a polynomial of degree n is considered in the following form:

$$\begin{aligned} g(x) &= p_n(x) = a_0 + a_1(x-x_0) + a_2(x-x_0)(x-x_1) + \cdots \\ &\quad + a_n(x-x_0)(x-x_1)\ldots(x-x_{n-1}). \end{aligned} \tag{19.160}$$

This is called the *Newton interpolation formula*, and it gives an easy calculation of the coefficients a_i $(i = 0,1,\ldots,n)$, since the interpolation conditions (19.159) result in a linear system of equations

with a triangular matrix.

■ For $n = 2$ one gets the annexed system of equations from (19.159). The interpolation polynomial $p_n(x)$ is uniquely determined by the interpolation conditions (19.159).

$$\begin{aligned} p_2(x_0) &= a_0 &&= y_0 \\ p_2(x_1) &= a_0 + a_1(x_1 - x_0) &&= y_1 \\ p_2(x_2) &= a_0 + a_1(x_2 - x_0) + a_2(x_2 - x_0)(x_2 - x_1) &&= y_2 \end{aligned}$$

The calculation of the function values can be simplified by the Horner schema (see 19.1.2.1, p. 952).

19.6.1.2 Lagrange's Interpolation Formula

A polynomial of n-th degree can be fitted through $n + 1$ points (x_ν, y_ν) $(\nu = 0, 1, \ldots, n)$, with the Lagrange formula:

$$g(x) = p_n(x) = \sum_{\mu=0}^{n} y_\mu L_\mu(x). \tag{19.161}$$

Here $L_\mu(x)$ $(\mu = 0, 1, \ldots, n)$ are the Lagrange interpolation polynomials. Equation (19.161) satisfies the interpolation conditions (19.159), since

$$L_\mu(x_\nu) = \delta_{\mu\nu} = \begin{cases} 1 & \text{for } \mu = \nu, \\ 0 & \text{for } \mu \neq \nu. \end{cases} \tag{19.162}$$

Here $\delta_{\mu\nu}$ is the Kronecker symbol. The Lagrange interpolation polynomials are defined by the formula

$$L_\mu = \frac{(x - x_0)(x - x_1)\cdots(x - x_{\mu-1})(x - x_{\mu+1})\cdots(x - x_n)}{(x_\mu - x_0)(x_\mu - x_1)\cdots(x_\mu - x_{\mu-1})(x_\mu - x_{\mu+1})\cdots(x_\mu - x_n)} = \prod_{\substack{\nu=0 \\ \nu\neq\mu}}^{n} \frac{x - x_\nu}{x_\mu - x_\nu}. \tag{19.163}$$

■ A polynomial is fitted through the points given by the table

x	0	1	3
y	1	3	2

The Lagrange interpolation formula (19.161) is used:

$$\begin{aligned} L_0(x) &= \frac{(x-1)(x-3)}{(0-1)(0-3)} = \frac{1}{3}(x-1)(x-3), \\ L_1(x) &= \frac{(x-0)(x-3)}{(1-0)(1-3)} = -\frac{1}{2}x(x-3), \\ L_2(x) &= \frac{(x-0)(x-1)}{(3-0)(3-1)} = \frac{1}{6}x(x-1); \end{aligned}$$

$$p_2(x) = 1 \cdot L_0(x) + 3 \cdot L_1(x) + 2 \cdot L_2(x) = -\frac{5}{6}x^2 + \frac{17}{6}x + 1.$$

The Lagrange interpolation formula depends explicitly and linearly on the given values y_μ of the function. This is its theoretical importance (see, e.g., the rule of Adams-Bashforth, 19.4.1.3, **3.**, p. 971). For practical calculation the Lagrange interpolation formula is rarely reasonable.

19.6.1.3 Aitken-Neville Interpolation

In several practical cases, the explicit form of the polynomial $p_n(x)$ is not needed, but only its value at a given location x of the interpolation domain. These function values can be obtained in a recursive way developed by Aitken and Neville. The useful notation

$$p_n(x) = p_{0,1,\ldots,n}(x), \tag{19.164}$$

is applied in which the interpolation points $x_0, x_1, \ldots, x_n$ and the degree n of the polynomial are denoted. Notice that

$$p_{0,1,\ldots,n}(x) = \frac{(x - x_0)p_{1,2,\ldots,n}(x) - (x - x_n)p_{0,1,2,\ldots,n-1}(x)}{x_n - x_0}, \tag{19.165}$$

i.e., the function value $p_{0,1,\ldots,n}(x)$ can be obtained by linear interpolation of the function values of $p_{1,2,\ldots,n}(x)$ and $p_{0,1,2,\ldots,n-1}(x)$, two interpolation polynomials of degree $\le n-1$. Application of (19.165) leads to a scheme which is given here for the case of $n = 4$:

$$\begin{array}{c|l} x_0 & y_0 = p_0 \\ x_1 & y_1 = p_1 \; p_{01} \\ x_2 & y_2 = p_2 \; p_{12} \; p_{012} \\ x_3 & y_3 = p_3 \; p_{23} \; p_{123} \; p_{0123} \\ x_4 & y_4 = p_4 \; p_{34} \; p_{234} \; p_{1234} \; \underline{p_{01234} = p_4(x).} \end{array} \qquad (19.166)$$

The elements of (19.166) are calculated column-wise. A new value in the scheme is obtained from its west and north-west neighbors

$$p_{23} = \frac{(x-x_2)p_3 - (x-x_3)p_2}{x_3 - x_2} = p_3 + \frac{x - x_3}{x_3 - x_2}(p_3 - p_2), \qquad (19.167a)$$

$$p_{123} = \frac{(x-x_1)p_{23} - (x-x_3)p_{12}}{x_3 - x_1} = p_{23} + \frac{x - x_3}{x_3 - x_1}(p_{23} - p_{12}), \qquad (19.167b)$$

$$p_{1234} = \frac{(x-x_1)p_{234} - (x-x_4)p_{123}}{x_4 - x_1} = p_{234} + \frac{x - x_4}{x_4 - x_1}(p_{234} - p_{123}). \qquad (19.167c)$$

For performing the *Aitken-Neville algorithm* on a computer only a vector $\underline{\mathbf{p}}$ with $n+1$ components (see [19.4]) is introduced, which takes the values of the columns in (19.166) after each other according to the rule that the value $p_{i-k,i-k+1,\ldots,i}$ $(i = k, k+1, \ldots, n)$ of the k-th column will be the i-th component p_i of $\underline{\mathbf{p}}$. The columns of (19.166) must be calculated from down to the top, so p contains all necessary values. The algorithm has the following two steps:

1. For $i = 0, 1, \ldots, n$ set $p_i = y_i$. (19.168a)

2. For $k = 1, 2, \ldots, n$ and for $i = n, n-1, \ldots, k$ compute $p_i = p_i + \dfrac{x - x_i}{x_i - x_{i-k}}(p_i - p_{i-1})$. (19.168b)

After finishing (19.168b) we have the required function value $p_n(x)$ at x in element p_n.

19.6.2 Approximation in Mean

The principle of approximation in mean is known as the *Gauss least squares method*. In calculations continuous and discrete cases are distinguished.

19.6.2.1 Continuous Problems, Normal Equations

The function $f(x)$ is approximated by a function $g(x)$ on the interval $[a, b]$ so that the expression

$$F = \int_a^b \omega(x)[f(x) - g(x)]^2 \, dx, \qquad (19.169)$$

depending on the parameters contained by $g(x)$, should be minimal. $\omega(x)$ denotes a given weight function, such that $\omega(x) > 0$ in the integration interval.
If the best approximation $g(x)$ is supposed to have the form

$$g(x) = \sum_{i=0}^{n} a_i g_i(x) \qquad (19.170)$$

with suitable linearly independent functions $g_0(x)$, $g_1(x)$, ... , $g_n(x)$, then the necessary conditions

$$\frac{\partial F}{\partial a_i} = 0 \quad (i = 0, 1, \ldots, n) \qquad (19.171)$$

for an extreme value of (19.169) result in the so-called *normal system of equations*

$$\sum_{i=0}^{n} a_i(g_i,\, g_k) = (f, g_k) \quad (k = 0, 1, \ldots, n) \tag{19.172}$$

to determine the unknown coefficients a_i. Here the brief notations

$$(g_i, g_k) = \int_a^b \omega(x) g_i(x) g_k(x)\, dx, \tag{19.173a}$$

$$(f, g_k) = \int_a^b \omega(x) f(x) g_k(x)\, dx \quad (i, k = 0, 1, \ldots, n) \tag{19.173b}$$

are used, which are considered as the *scalar products* of the two indicated functions.
The system of normal equations can be solved uniquely, since the functions $g_0(x),\ g_1(x),\ \ldots,\ g_n(x)$ are linearly independent. The coefficient matrix of the system (19.172) is symmetric, so the Cholesky method (see 19.2.1.2, p. 958) can be applied. The coefficients a_i can be determined directly, without solving the system of equations, if the system of functions $g_i(x)$ is *orthogonal*, that is, if

$$(g_i, g_k) = 0 \ \text{ for } \ i \neq k. \tag{19.174}$$

It is called an *orthonormal* system, if

$$(g_i, g_k) = \begin{cases} 0 & \text{for } i \neq k, \\ 1 & \text{for } i = k \end{cases} \quad (i, k = 0, 1, \ldots, n). \tag{19.175}$$

With (19.175), the normal equations (19.172) are reduced to

$$a_i = (f, g_i) \quad (i = 0, 1, \ldots, n). \tag{19.176}$$

Linearly independent function systems can be orthogonalized. From the power functions $g_i(x) = x^i$ $(i = 0, 1, \ldots, n)$, depending on the weight function and on the interval, the *orthogonal polynomials* in **Table 19.2** can be obtained.

Table 19.2 Orthogonal polynomials

$[a, b]$	$\omega(x)$	Name of the polynomials	see p.
$[-1, 1]$	1	Legendre polynomial $P_n(x)$	566
$[-1, 1]$	$\dfrac{1}{\sqrt{1-x^2}}$	Chebysev polynomial $T_n(x)$	989
$[0, \infty)$	e^{-x}	Laguerre polynomial $L_n(x)$	568
$(-\infty, \infty)$	$e^{-x^2/2}$	Hermite polynomial $H_n(x)$	568

(19.177)

These polynomial systems can be used on arbitrary intervals:
1. Finite approximation interval.
2. Approximation interval infinite at one end, e.g., in time-dependent problems.
3. Approximation interval infinite at both ends, e.g., in stream problems.
Every finite interval $[a, b]$ can be transformed by the substitution

$$x = \frac{b+a}{2} + \frac{b-a}{2} t \quad (x \in [a, b],\ t \in [-1, 1]) \tag{19.178}$$

into the interval $[-1, 1]$.

19.6.2.2 Discrete Problems, Normal Equations, Householder's Method

Let N pairs of values (x_ν, y_ν) be given, e.g., by measured values. A function $g(x)$ has to be determined, whose values $g(x_\nu)$ differ from the given values y_ν in such a way that the quadratic expression

$$F = \sum_{\nu=1}^{N} [y_\nu - g(x_\nu)]^2 \tag{19.179}$$

is minimal. The value of F depends on the parameters contained in the function $g(x)$. Formula (19.179) represents the classical *sum of residual squares*. The minimization of the sum of residual squares is called the *least squares method*. From the assumption (19.170) and the necessary conditions $\frac{\partial F}{\partial a_i} = 0$ $(i = 0, 1, \ldots, n)$ for a relative minimum of (19.179) for the coefficients the *normal system of equations* is obtained:

$$\sum_{i=0}^{n} a_i[g_i g_k] = [y g_k] \quad (k = 0, 1, \ldots, n). \tag{19.180}$$

Here the Gaussean sum symbols are used in the following notation:

$$[g_i g_k] = \sum_{\nu=1}^{N} g_i(x_\nu) g_k(x_\nu), \tag{19.181a}$$

$$[y g_k] = \sum_{\nu=1}^{N} y_\nu g_k(x_\nu) \quad (i, k = 0, 1, \ldots, n). \tag{19.181b}$$

Usually, $n \ll N$.

■ For the polynomial $g(x) = a_0 + a_1 x + \cdots + a_n x^n$, the normal equations are $a_0[x^k] + a_1[x^{k+1}] + \cdots + a_n[x^{k+n}] = [x^k y]$ $(k = 0, 1, \ldots, n)$ with $[x^k] = \sum_{\nu=1}^{N} {x_\nu}^k$, $[x^0] = N$, $[x^k y] = \sum_{\nu=1}^{N} {x_\nu}^k y_\nu$, $[y] = \sum_{\nu=1}^{N} y_\nu$. The coefficient matrix of the normal system of equations (19.180) is symmetric, so for the numerical solution the Cholesky method can be applied.

The normal equations (19.180) and the residue sum square (19.179) have the following compact form:

$$\mathbf{G}^{\mathrm{T}}\mathbf{G}\underline{\mathbf{a}} = \mathbf{G}^{\mathrm{T}}\underline{\mathbf{y}}, \quad F = (\underline{\mathbf{y}} - \mathbf{G}\underline{\mathbf{a}})^{\mathrm{T}}(\underline{\mathbf{y}} - \mathbf{G}\underline{\mathbf{a}}) \quad \text{with} \tag{19.182a}$$

$$\mathbf{G} = \begin{pmatrix} g_0(x_1) & g_1(x_1) & g_2(x_1) & \ldots & g_n(x_1) \\ g_0(x_2) & g_1(x_2) & g_2(x_2) & \ldots & g_n(x_2) \\ g_0(x_3) & g_1(x_3) & g_2(x_3) & \ldots & g_n(x_3) \\ \vdots & & & & \\ g_0(x_N) & g_1(x_N) & g_2(x_N) & \ldots & g_n(x_N) \end{pmatrix}, \quad \underline{\mathbf{y}} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_N \end{pmatrix}, \quad \underline{\mathbf{a}} = \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix}. \tag{19.182b}$$

If, instead of the minimalization of the sum of residual squares, one wants to solve the interpolation problem for the N points (x_ν, y_ν), then the following system of equations should be solved:

$$\mathbf{G}\underline{\mathbf{a}} = \underline{\mathbf{y}}. \tag{19.183}$$

This system of equations is over-determined in the case of $n < N - 1$, and usually it does not have any solution. The equations (19.180) or (19.182a) are obtained by multiplying (19.183) by $\mathbf{G}^{\mathrm{T}}$.

From a numerical viewpoint, the Householder method (see 4.5.3.2, **2.**, p. 314) is recommended to solve equation (19.183), and this solution results in the minimal sum of residual squares (19.179).

19.6.2.3 Multidimensional Problems

1. Computation of Adjustments

Suppose that there is a function $f(x_1, x_2, \ldots, x_n)$ of n independent variables $x_1, x_2, \ldots, x_n$. Its explicit form is not known; only N substitution values f_ν are given, which are, in general, measured values. These data can be written in a table (see (19.184)).

$$\begin{array}{c|cccc} x_1 & x_1^{(1)} & x_1^{(2)} & \ldots & x_1^{(N)} \\ x_2 & x_2^{(1)} & x_2^{(2)} & \ldots & x_2^{(N)} \\ \vdots & \vdots & \vdots & & \vdots \\ x_n & x_n^{(1)} & x_n^{(2)} & \ldots & x_n^{(N)} \\ \hline f & f_1 & f_2 & \ldots & f_N \end{array} \tag{19.184}$$

The formulation of the adjustment problem is clearer by introducing the following vectors:

$\underline{\mathbf{x}} = (x_1, x_2, \dots, x_n)^{\mathrm{T}}$: Vector of n independent variables,
$\underline{\mathbf{x}}^{(\nu)} = (x_1^{(\nu)}, x_2^{(\nu)}, \dots, x_n^{(\nu)})^{\mathrm{T}}$: Vector of the ν-th interpolation node $(\nu = 1, \dots, N)$,
$\underline{\mathbf{f}} = (f_1, f_2, \dots f_N)^{\mathrm{T}}$: Vector of the N function values at the N interpolation nodes.

$f(x_1, x_2, \dots, x_n) = f(\underline{\mathbf{x}})$ is approximated by a function of the form

$$g(x_1, x_2, \dots, x_n) = \sum_{i=0}^{m} a_i g_i(x_1, x_2, \dots, x_n). \qquad (19.185)$$

Here, the $m+1$ functions $g_i(x_1, x_2, \dots, x_n) = g_i(\underline{\mathbf{x}})$ are suitable, selected functions.

■ **A:** Linear approximation by n variables: $g(x_1, x_2, \dots, x_n) = a_0 + a_1x_1 + a_2x_2 + \dots + a_nx_n$.

■ **B:** Complete quadratic approximation with three variables:
$g(x_1, x_2, x_3) = a_0 + a_1x_1 + a_2x_2 + a_3x_3 + a_4{x_1}^2 + a_5{x_2}^2 + a_6{x_3}^2 + a_7x_1x_2 + a_8x_1x_3 + a_9x_2x_3$.

The coefficients are chosen to minimize $\sum_{\nu=1}^{N} \left[f_\nu - g\left(x_1^{(\nu)}, x_2^{(\nu)}, \dots, x_n^{(\nu)}\right)\right]^2$.

2. Normal System of Equations

Analogously to (19.182b) the matrix $\mathbf{G}$ is formed in which the interpolation nodes x_ν are replaced by vectorial interpolation nodes $\underline{\mathbf{x}}^{(\nu)}$ $(\nu = 1, 2, \dots, N)$. To determine the coefficients, the normal system of equations

$$\mathbf{G}^{\mathrm{T}}\mathbf{G}\underline{\mathbf{a}} = \mathbf{G}^{\mathrm{T}}\underline{\mathbf{f}} \qquad (19.186)$$

can be used or the over-determined system of equations

$$\mathbf{G}\underline{\mathbf{a}} = \underline{\mathbf{f}}. \qquad (19.187)$$

■ For an example of multidimensional regression see 16.3.4.3, **3.**, p. 842.

19.6.2.4 Non-Linear Least Squares Problems

The main idea is discussed for a one-dimensional discrete case. The approximation function $g(x)$ depends non-linearly on certain parameters.

■ **A:** $g(x) = a_0e^{a_1x} + a_2e^{a_3x}$. This expression does not depend linearly on the parameters a_1 and a_3.

■ **B:** $g(x) = a_0e^{a_1x}\cos a_2x$. This function does not depend linearly on the parameters a_1 and a_2.

The fact that the approximation function $g(x)$ depends on a parameter vector $\underline{\mathbf{a}} = (a_0, a_1, \dots, a_n)^{\mathrm{T}}$ is indicated by the notation

$$g = g(x, \underline{\mathbf{a}}) = g(x;\ a_0, a_1, \dots, a_n). \qquad (19.188)$$

Suppose, N pairs of values (x_ν, y_ν) $(\nu = 1, 2, \dots, N)$ are given. To minimize the sum of residual squares

$$\sum_{\nu=1}^{N} [y_\nu - g(x_\nu;\ a_0, a_1, \dots, a_n)]^2 = F(a_0, a_1, \dots, a_n) \qquad (19.189)$$

the necessary conditions $\dfrac{\partial F}{\partial a_i} = 0$ $(i = 0, 1, \dots, n)$ lead to a non-linear normal equation system which must be solved by an iterative method, e.g., by the Newton method (see 19.2.2.2, p. 962).
Another way to solve the problem, which is usually used in practical problems, is the application of the Gauss-Newton method (see 19.2.2.3, p. 962) given for the solution of the non-linear least squares problem (19.24). The following steps are needed to apply it for this non-linear approximation problem (19.189):

1. Linearization of the approximating function $g(x, \underline{\mathbf{a}})$ with the help of the Taylor formula with respect to a_i. To do this, the approximation values $a_i^{(0)}$ $(i = 0, 1, \dots, n)$ are needed:

$$g(x, \underline{\mathbf{a}}) \approx \tilde{g}(x, \underline{\mathbf{a}}) = g(x, \underline{\mathbf{a}}^{(0)}) + \sum_{i=0}^{n} \frac{\partial g}{\partial a_i}(x, \underline{\mathbf{a}}^{(0)})(a_i - a_i^{(0)}). \qquad (19.190)$$

2. Solution of the linear minimum problem

$$\sum_{\nu=1}^{N}[y_\nu - \tilde{g}(x_\nu, \underline{\mathbf{a}})]^2 = \text{min!} \tag{19.191}$$

with the help of the normal equation system

$$\tilde{\mathbf{G}}^{\mathrm{T}}\tilde{\mathbf{G}}\underline{\boldsymbol{\Delta}\mathbf{a}} = \tilde{\mathbf{G}}^{\mathrm{T}}\underline{\boldsymbol{\Delta}\mathbf{y}} \tag{19.192}$$

or by the Householder method. In (19.192) the components of the vectors $\underline{\boldsymbol{\Delta}\mathbf{a}}$ and $\underline{\boldsymbol{\Delta}\mathbf{y}}$ are given as

$$\Delta a_i = a_i - a_i^{(0)} \qquad (i = 0, 1, 2, \dots, n) \text{ and} \tag{19.193a}$$

$$\Delta y_\nu = y_\nu - g(x_\nu, \underline{\mathbf{a}}^{(0)}) \quad (\nu = 1, 2, \dots, N). \tag{19.193b}$$

The matrix $\tilde{G}$ can be determined analogously to G in (19.182b), where $g_i(x_\nu)$ are replaced by $\dfrac{\partial g}{\partial a_i}(x_\nu, \underline{\mathbf{a}}_1^{(0)})$ $(i = 0, 1, \dots, n;\ \nu = 1, 2, \dots, N)$.

3. Calculation of a new approximation

$$a_i^{(1)} = a_i^{(0)} + \Delta a_i \text{ or } a_i^{(1)} = a_i^{(0)} + \gamma\Delta a_i \quad (i = 0, 1, 2, \dots, n), \tag{19.194}$$

where $\gamma > 0$ is a step length parameter.

By repeating steps **2** and **3** with $a_i^{(1)}$ instead of $a_i^{(0)}$, etc. a sequence of approximation values is obtained for the required parameters, whose convergence strongly depends on the accuracy of the initial approximations. The value of the sum of residual squares can be reduced with the introduction of the multiplier γ.

19.6.3 Chebyshev Approximation

19.6.3.1 Problem Definition and the Alternating Point Theorem

1. Principle of Chebyshev Approximation

Chebyshev approximation or *uniform approximation* in the continuous case is the following: The function $f(x)$ has to be approximated in an interval $a \le x \le b$ by the approximation function $g(x) = g(x;\ a_0, a_1, \dots, a_n)$ so that the error defined by

$$\max_{a \le x \le b} |f(x) - g(x;\ a_0, a_1, \dots, a_n)| = \Phi(a_0, a_1, \dots, a_n) \tag{19.195}$$

should be as small as possible for the appropriate choice of the unknown parameters a_i $(i = 0, 1, \dots, n)$. If there exists such an approximating function for $f(x)$, then the maximum of the absolute error value will be taken at least at $n+2$ points x_ν of the interval, at the so-called *alternating points*, with changing signs (**Fig. 19.10**). This is actually the meaning of the *alternating point theorem* for the characterization of the solution of a Chebyshev approximation problem.

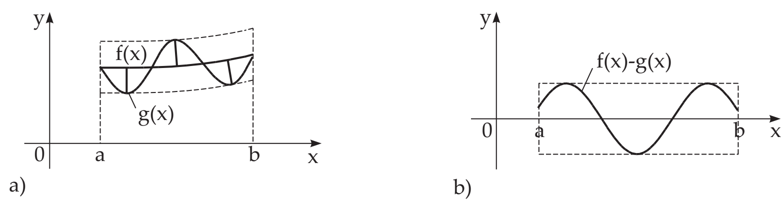


Figure 19.10

■ If the function $f(x) = x^n$ is approximated on the interval $[-1, 1]$ by a polynomial of degree $\le n-1$ in the Chebyshev sense, then the *Chebyshev polynomial* $T_n(x)$ is obtained as an error function whose

maximum is normed to one. The alternating points, being at the endpoints and at exactly $n-1$ points in the interior of the interval, correspond to the extreme points of $T_n(x)$ (**Fig. 19.11a–f**).

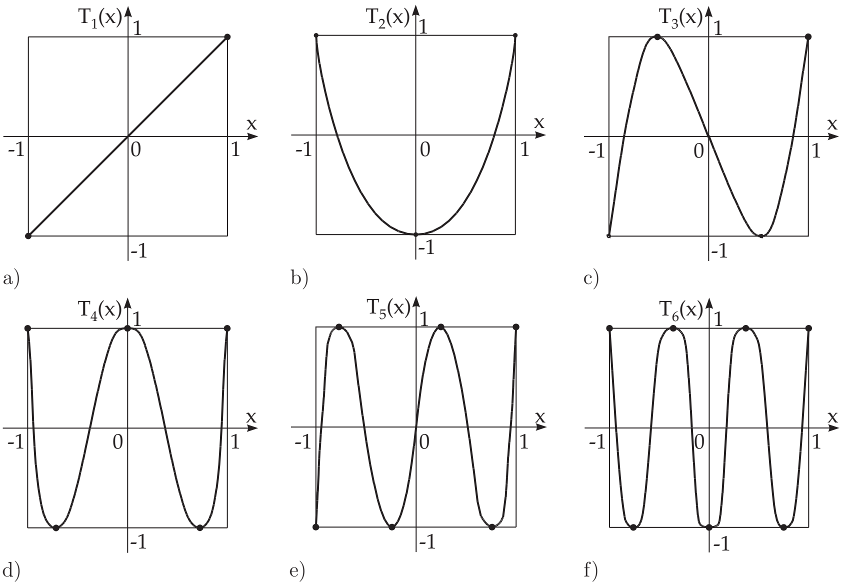


Figure 19.11

19.6.3.2 Properties of the Chebyshev Polynomials

1. Representation

$$T_n(x) = \cos(n \arccos x), \tag{19.196a}$$

$$T_n(x) = \frac{1}{2}\left[\left(x+\sqrt{x^2-1}\right)^n + \left(x-\sqrt{x^2-1}\right)^n\right], \tag{19.196b}$$

$$T_n(x) = \begin{cases} \cos nt,\ x = \cos t & \text{for } |x| < 1, \\ \cosh nt,\ x = \cosh t & \text{for } |x| > 1 \end{cases} \quad (n = 1, 2, \ldots). \tag{19.196c}$$

2. Roots of $T_n(x)$

$$x_\mu = \cos\frac{(2\mu-1)\pi}{2n} \quad (\mu = 1, 2, \ldots, n). \tag{19.197}$$

3. Position of the extreme values of $T_n(x)$ for $x \in [-1, 1]$

$$x_\nu = \cos\frac{\nu\pi}{n} \quad (\nu = 0, 1, 2, \ldots, n). \tag{19.198}$$

4. Recursion Formula

$$T_{n+1} = 2xT_n(x) - T_{n-1}(x) \quad (n = 1, 2, \ldots;\ T_0(x) = 1,\ T_1(x) = x). \tag{19.199}$$

This recursion results in e.g.

$$T_2(x) = 2x^2 - 1, \quad T_3(x) = 4x^3 - 3x, \tag{19.200a}$$

$$T_4(x) = 8x^4 - 8x^2 + 1, \quad T_5(x) = 16x^5 - 20x^3 + 5x, \tag{19.200b}$$

$$T_6(x) = 32x^6 - 48x^4 + 18x^2 - 1, \tag{19.200c}$$

$$T_7(x) = 64x^7 - 112x^5 + 56x^3 - 7x, \tag{19.200d}$$

$$T_8(x) = 128x^8 - 256x^6 + 160x^4 - 32x^2 + 1, \tag{19.200e}$$

$$T_9(x) = 256x^9 - 576x^7 + 432x^5 - 120x^3 + 9x, \tag{19.200f}$$

$$T_{10}(x) = 512x^{10} - 1280x^8 + 1120x^6 - 400x^4 + 50x^2 - 1. \tag{19.200g}$$

19.6.3.3 Remes Algorithm

1. Consequences of the Alternating Point Theorem

The numerical solution of the continuous Chebyshev approximation problem originates from the alternating point theorem. The approximating function is chosen as

$$g(x) = \sum_{i=0}^{n} a_i g_i(x) \tag{19.201}$$

with $n+1$ linearly independent known functions, and the coefficients of the solution of the Chebyshev problem are denoted by $a_i{}^*$ $(i = 0, 1, \ldots, n)$ and the minimal deviation according to (19.195) by $\varrho = \Phi(a_0{}^*, a_1{}^*, \ldots, a_n{}^*)$. In the case when the functions f and g_i $(i = 0, 1, \ldots, n)$ are differentiable, from the alternating point theorem one has

$$\sum_{i=0}^{n} a_i{}^* g_i(x_\nu) + (-1)^\nu \varrho = f(x_\nu), \quad \sum_{i=0}^{n} a_i{}^* g_i'(x_\nu) = f'(x_\nu) \quad (\nu = 1, 2, \ldots, n+2). \tag{19.202}$$

The nodes x_ν are the alternating points with

$$a \le x_1 < x_2 < \ldots < x_{n+2} \le b. \tag{19.203}$$

The equations (19.202) give $2n + 4$ conditions for the $2n + 4$ unknown quantities of the Chebyshev approximation problem: $n + 1$ coefficients, $n + 2$ alternating points and the minimal deviation ϱ. If the endpoints of the interval belong to the alternating points, then the conditions for the derivatives are not necessarily valid there.

2. Determination of the Minimal Solution according to Remes

According to Remes, one proceeds with the numerical determination of the minimal solution as follows:

1. An approximation of the alternating points $x_\nu{}^{(0)}$ $(\nu = 1, 2, \ldots, n+2)$ are determined according to (19.203), e.g., equidistant or as the positions of the extrema of $T_{n+1}(x)$ (see 19.6.3.2, p. 988).

2. The linear system of equations

$$\sum_{i=0}^{n} a_i g_i(x_\nu{}^{(0)}) + (-1)^\nu \varrho = f(x_\nu{}^{(0)}) \quad (\nu = 1, 2, \ldots, n+2)$$

is solved and the solutions are the approximations $a_i{}^{(0)}$ $(i = 0, 1, \ldots, n)$ and ϱ_0.

3. A new approximation of the alternating points $x_\nu{}^{(1)}$ $(\nu = 1, 2, \ldots, n+2)$ is determined, e.g., as positions of the extrema of the error function $f(x) - \sum\limits_{i=0}^{n} a_i{}^{(0)} g_i(x)$. Now, it is sufficient to apply only approximations of these points.

By repeating steps **2** and **3** with $x_\nu{}^{(1)}$ and $a_i{}^{(1)}$ instead of $x_\nu{}^{(0)}$ and $a_i{}^{(0)}$, etc. a sequence of approximations is obtained for the coefficients and the alternating points, whose convergence is guaranteed under certain conditions, which can be given (see [19.25]). The calculations are stopped if, e.g., from a certain iteration index μ

$$|\varrho_\mu| = \max_{a \le x \le b} \left| f(x) - \sum_{i=0}^{n} a_i{}^{(\mu)} g_i(x) \right| \tag{19.204}$$

holds with a sufficient accuracy.

19.6.3.4 Discrete Chebyshev Approximation and Optimization

From the continuous Chebyshev approximation problem

$$\max_{a\le x\le b}\left|f(x)-\sum_{i=0}^{n}a_i g_i(x)\right| = \min! \qquad (19.205)$$

the corresponding discrete problem can be got, if requiring N nodes x_ν $(\nu = 1,2,\ldots,N;\ N \ge n+2)$ are chosen with the property $a \le x_1 < x_2 \cdots < x_N \le b$ and requiring

$$\max_{\nu=1,2,\ldots,N}\left|f(x_\nu)-\sum_{i=0}^{n}a_i g_i(x_\nu)\right| = \min!\,. \qquad (19.206)$$

The substitution

$$\gamma = \max_{\nu=1,2,\ldots,N}\left|f(x_\nu)-\sum_{i=0}^{n}a_i g_i(x_\nu)\right|, \qquad (19.207)$$

has obviously the consequence

$$\left|f(x_\nu)-\sum_{i=0}^{n}a_i g_i(x_\nu)\right| \le \gamma \quad (\nu = 1,2,\ldots,N). \qquad (19.208)$$

Eliminating the absolute values from (19.208) a linear system of inequalities is obtained for the coefficients a_i and γ, so the problem (19.206) becomes a linear programming problem (see 18.1.1.1, p. 909):

$$\gamma = \min! \quad \text{subject to} \quad \begin{cases} \gamma + \sum\limits_{i=0}^{n} a_i g_i(x_\nu) \ge f(x_\nu), \\ \gamma - \sum\limits_{i=0}^{n} a_i g_i(x_\nu) \ge -f(x_\nu) \end{cases} \quad (\nu = 1,2,\ldots,N). \qquad (19.209)$$

Equation (19.209) has a minimal solution with $\gamma > 0$. For a sufficiently large number N of nodes and with some further conditions the solution of the discrete problem can be considered as the solution of the continuous problem.

If instead of the linear approximation function $g(x) = \sum\limits_{i=0}^{n} a_i g_i(x)$ a non-linear approximation function $g(x) = g(x; a_0, a_1, \ldots, a_n)$ is used, which does not depend linearly on the parameters $a_0, a_1, \ldots, a_n$, then analogously a *non-linear optimization problem* is obtained. It is usually non-convex even in the cases of simple function forms. This essentially reduces the number of numerical solution methods for non-linear optimization problems (see 18.2.2.1, p. 926).

19.6.4 Harmonic Analysis

A periodic function $f(x)$ with period 2π, which is given formally or empirically, should be approximated by a *trigonometric polynomial* or a *Fourier sum* of the form

$$g(x) = \frac{a_0}{2} + \sum_{k=1}^{n} (a_k \cos kx + b_k \sin kx), \tag{19.210}$$

where the coefficients a_0, a_k and b_k are unknown real numbers. The determination of the coefficients is the topic of harmonic analysis.

19.6.4.1 Formulas for Trigonometric Interpolation

1. Formulas for the Fourier Coefficients

Since the function system 1, $\cos kx$, $\sin kx$ $(k = 1, 2, \ldots, n)$ is orthogonal in the interval $[0, 2\pi]$ with respect to the weight function $\omega \equiv 1$, the formulas for the coefficients are obtained as

$$a_k = \frac{1}{\pi} \int_0^{2\pi} f(x) \cos kx \, dx, \quad b_k = \frac{1}{\pi} \int_0^{2\pi} f(x) \sin kx \, dx \quad (k = 0, 1, 2, \ldots, n) \tag{19.211}$$

by applying the continuous least squares method according to (19.172). The coefficients a_k and b_k calculated by formulas (19.211) are called *Fourier coefficients* of the periodic function $f(x)$ (see 7.4, p. 474).

If the integrals in (19.211) are complicated or the function $f(x)$ is known only at discrete points, then the Fourier coefficients can be determined only approximately by numerical integration.

Using the trapezoidal formula (see 19.3.2.2, p. 964) with $N + 1$ equidistant nodes

$$x_\nu = \nu h \quad (\nu = 0, 1, \ldots, N), \quad h = \frac{2\pi}{N} \tag{19.212}$$

the approximation formula

$$a_k \approx \tilde{a}_k = \frac{2}{N} \sum_{\nu=1}^{N} f(x_\nu) \cos kx_\nu, \quad b_k \approx \tilde{b}_k = \frac{2}{N} \sum_{\nu=1}^{N} f(x_\nu) \sin kx_\nu \quad (k = 0, 1, 2, \ldots, n) \tag{19.213}$$

is obtained. The trapezoidal formula becomes the very simple rectangular formula in the case of periodic functions. It has higher accuracy here as a consequence of the following fact: If $f(x)$ is periodic and $(2m + 2)$ times differentiable, then the trapezoidal formula has an error of order $O(h^{2m+2})$.

2. Trigonometric Interpolation

Some special trigonometric polynomials formed with the approximation coefficients $\tilde{a}_k$ and $\tilde{b}_k$ have important properties. Two of them are mentioned here:

1. Interpolation Suppose $N = 2n$ holds. The special trigonometric polynomial

$$\tilde{g}_1(x) = \frac{1}{2}\tilde{a}_0 + \sum_{k=1}^{n-1} (\tilde{a}_k \cos kx + \tilde{b}_k \sin kx) + \frac{1}{2}\tilde{a}_n \cos nx \tag{19.214}$$

with coefficients (19.213) satisfies the interpolation conditions

$$\tilde{g}_1(x_\nu) = f(x_\nu) \qquad (\nu = 1, 2, \ldots, N) \tag{19.215}$$

at the interpolation nodes x_ν (19.212). Because of the perodicity of $f(x)$ $f(x_0) = f(x_N)$ holds.

2. Approximation in Mean Suppose $N = 2n$. The special trigonometric polynomial

$$\tilde{g}_2(x) = \frac{1}{2}\tilde{a}_0 + \sum_{k=1}^{m} (\tilde{a}_k \cos kx + \tilde{b}_k \sin kx) \tag{19.216}$$

for $m < n$ and with the coefficients (19.213) approximates the function $f(x)$ in discrete quadratic mean with respect to the N nodes x_ν (19.212), that is, the residual sum of squares

$$F = \sum_{\nu=1}^{N} [f(x_\nu) - \tilde{g}_2(x_\nu)]^2 \tag{19.217}$$

is minimal. The formulas (19.213) are the originating point for the different ways of effective calculation of Fourier coefficients.

19.6.4.2 Fast Fourier Transformation (FFT)

1. Computation costs of computing Fourier coefficients

The sums in the formulas (19.213) also occur in connection with discrete Fourier transformation, e.g., in electrotechnics, in impulse and picture processing. Here N can be very large, so the occurring sums must be calculated in a rational way, since the calculation of the N approximating values (19.213) of the Fourier coefficients requires about N^2 additions and multiplications.

For the special case of $N = 2^p$, the number of multiplications can be largely reduced from $N^2(= 2^{2p})$ to $pN(= p2^p)$ with the help of the so-called *fast Fourier transformation* **FFT**. The magnitude of this reduction is demonstrated on the example on the right-hand side.

p	N^2	pN
10	$\sim 10^6$	$\sim 10^4$
20	$\sim 10^{12}$	$\sim 10^7$

(19.218)

By this method, the computation costs and computation time are reduced so effectively that in some important application fields even a smaller computer is sufficient.

The FFT uses the properties of the N-th unit roots, i.e., the solutions of equation $z^N = 1$ to a successive sum up in (19.213).

2. Complex Representation of the Fourier Sum

The principle of FFT can be described fairly easily if the Fourier sum (19.210) is rewritten with the formulas

$$\cos kx = \frac{1}{2}\left(e^{ikx} + e^{-ikx}\right), \quad \sin kx = \frac{i}{2}\left(e^{-ikx} - e^{ikx}\right) \tag{19.219}$$

into the complex form

$$g(x) = \frac{1}{2}a_0 + \sum_{k=1}^{n}(a_k \cos kx + b_k \sin kx) = \frac{1}{2}a_0 + \sum_{k=1}^{n}\left(\frac{a_k - ib_k}{2}e^{ikx} + \frac{a_k + ib_k}{2}e^{-ikx}\right). \tag{19.220}$$

By substitution

$$c_k = \frac{a_k - ib_k}{2}, \quad \text{(19.221a)} \qquad \text{because of (19.211)} \quad c_k = \frac{1}{2\pi}\int_0^{2\pi} f(x)e^{-ikx}\,dx, \tag{19.221b}$$

so (19.220) becomes the *complex representation of the Fourier sum*:

$$g(x) = \sum_{k=-n}^{n} c_k e^{ikx} \quad \text{with} \quad c_{-k} = \bar{c}_k. \tag{19.222}$$

If the complex coefficients c_k are known, then the required real Fourier coefficients can be got in the following simple way:

$$a_0 = 2c_0, \quad a_k = 2\mathrm{Re}(c_k), \quad b_k = -2\mathrm{Im}(c_k) \quad (k = 1, 2, \ldots, n). \tag{19.223}$$

3. Numerical Calculation of the Complex Fourier Coefficients

For the numerical determination of c_k the trapezoidal formula can be applied for (19.221b) analogously to (19.212) and (19.213), and the discrete complex Fourier coefficients $\tilde{c}_k$ are obtained:

$$\tilde{c}_k = \frac{1}{N}\sum_{\nu=0}^{N-1} f(x_\nu)e^{-ikx_\nu} = \sum_{\nu=0}^{N-1} f_\nu \omega_N^{k\nu} \quad (k = 0, 1, 2, \ldots, n) \quad \text{with} \tag{19.224a}$$

$$f_\nu = \frac{1}{N} f(x_\nu), \quad x_\nu = \frac{2\pi\nu}{N} \quad (\nu = 0, 1, 2, \ldots, N-1), \quad \omega_N = e^{-\frac{2\pi i}{N}}. \tag{19.224b}$$

Relation (19.224a) with the quantities (19.224b) is called the *discrete complex Fourier transformation* of length N of the values f_ν $(\nu = 0, 1, 2, \ldots, N-1)$.
The powers $\omega_N^\nu = z$ $(\nu = 0, 1, 2, \ldots, N-1)$ satisfy equation $z^N = 1$. So, they are called the *N-th unit roots*. Since $e^{-2\pi i} = 1$,

$$\omega_N^N = 1,\ \omega_N^{N+1} = \omega_N^1,\ \omega_N^{N+2} = \omega_N^2, \ldots . \tag{19.225}$$

The effective calculation of the sum (19.224a) uses the fact that a discrete complex Fourier transformation of length $N = 2n$ can be reduced to two transformations with length $\frac{N}{2} = n$ in the following way:

a) For every coefficient $\tilde{c}_k$ with an even index, i.e., $k = 2l$,

$$\tilde{c}_{2l} = \sum_{\nu=0}^{2n-1} f_\nu \omega_N^{2l\nu} = \sum_{\nu=0}^{n-1} \left[f_\nu \omega_N^{2l\nu} + f_{n+\nu} \omega_N^{2l(n+\nu)} \right] = \sum_{\nu=0}^{n-1} \left[f_\nu + f_{n+\nu} \right] \omega_N^{2l\nu} \tag{19.226}$$

holds. Here the equality $\omega_N^{2l(n+\nu)} = \omega_N^{2ln} \omega_N^{2l\nu} = \omega_N^{2l\nu}$ is used.
Substituting

$$y_\nu = f_\nu + f_{n+\nu} \quad (\nu = 0, 1, 2, \ldots, n-1) \tag{19.227}$$

and considering that $\omega_N^2 = \omega_n$, the sum

$$\tilde{c}_{2l} = \sum_{\nu=0}^{n-1} y_\nu \omega_n^{l\nu} \quad (\nu = 0, 1, 2, \ldots, n-1) \tag{19.228}$$

is the discrete complex Fourier transformation of the values y_ν $(\nu = 0, 1, 2, \ldots, n-1)$ with length $n = \frac{N}{2}$.

b) For every coefficient $\tilde{c}_k$ with an odd index, i.e., with $k = 2l + 1$

$$\tilde{c}_{2l+1} = \sum_{\nu=0}^{2n-1} f_\nu \omega_N^{(2l+1)\nu} = \sum_{\nu=0}^{n-1} \left[(f_\nu - f_{n+\nu}) \omega_N^\nu \right] \omega_N^{2l\nu} \tag{19.229}$$

is obtained analogously. Substituting

$$y_{n+\nu} = (f_\nu - f_{n+\nu}) \omega_N^\nu \quad (\nu = 0, 1, 2, \ldots, n-1) \tag{19.230}$$

and considering that $\omega_N^2 = \omega_n$, the sum

$$\tilde{c}_{2l+1} = \sum_{\nu=0}^{n-1} y_{n+\nu} \omega_n^{l\nu} \quad (\nu = 0, 1, 2, \ldots, n-1) \tag{19.231}$$

is the discrete complex Fourier transformation of the values $y_{n+\nu}$ $(\nu = 0, 1, 2, \ldots, n-1)$ with length $n = \frac{N}{2}$.
The reduction according to **a)** and **b)**, i.e., the reduction of a discrete complex Fourier transformation to two discrete complex Fourier transformations of half the length, can be continued if N is a power of 2, i.e., if $N = 2^p$ (p is a natural number). The application of the reduction after p times is called the FFT.
Since every reduction step requires $\frac{N}{2}$ complex multiplications because of (19.230), the computation cost of the FFT method is

$$\frac{N}{2} p = \frac{N}{2} \log_2 N. \tag{19.232}$$

4. Scheme for FFT

For the special case $N = 8 = 2^3$, the three corresponding reduction steps of the FFT according to (19.227) and (19.230) are demonstrated in the following **Scheme 1**:

Scheme 1:

	Step 1	Step 2	Step 3	
f_0	$y_0 = f_0 + f_4$	$y_0 := y_0 + y_2$	$y_0 := y_0 + y_1$	$= \tilde{c}_0$
f_1	$y_1 = f_1 + f_5$	$y_1 := y_1 + y_3$	$y_1 := (y_0 - y_1)\omega_2^0$	$= \tilde{c}_4$
f_2	$y_2 = f_2 + f_6$	$y_2 := (y_0 - y_2)\omega_4^0$	$y_2 := y_2 + y_3$	$= \tilde{c}_2$
f_3	$y_3 = f_3 + f_7$	$y_3 := (y_1 - y_3)\omega_4^1$	$y_3 := (y_2 - y_3)\omega_2^0$	$= \tilde{c}_6$
f_4	$y_4 = (f_0 - f_4)\omega_8^0$	$y_4 := y_4 + y_6$	$y_4 := y_4 + y_5$	$= \tilde{c}_1$
f_5	$y_5 = (f_1 - f_5)\omega_8^1$	$y_5 := y_5 + y_7$	$y_5 := (y_4 - y_5)\omega_2^0$	$= \tilde{c}_5$
f_6	$y_6 = (f_2 - f_6)\omega_8^2$	$y_6 := (y_4 - y_6)\omega_4^0$	$y_6 := y_6 + y_7$	$= \tilde{c}_3$
f_7	$y_7 = (f_3 - f_7)\omega_8^3$	$y_7 := (y_5 - y_7)\omega_4^1$	$y_7 := (y_6 - y_7)\omega_2^0$	$= \tilde{c}_7$
	$N = 8,\ n := 4,\ \omega_8 = \mathrm{e}^{-\frac{2\pi\mathrm{i}}{8}}$	$N := 4,\ n := 2,\ \omega_4 = \omega_8^2$	$N := 2,\ n := 1,\ \omega_2 = \omega_4^2$	

It can be observed how terms with even and odd indices appear. In **Scheme 2** (19.233) the structure of the method is illustrated.

Scheme 2:

$$\tilde{c}_k \Rightarrow \begin{cases} \tilde{c}_{2k} \quad \Rightarrow & \begin{cases} \tilde{c}_{4k} \quad \Rightarrow & \begin{cases} \tilde{c}_{8k} \\ \tilde{c}_{8k+4} \end{cases} \\ \tilde{c}_{4k+2} \quad \Rightarrow & \begin{cases} \tilde{c}_{8k+2} \\ \tilde{c}_{8k+6} \end{cases} \end{cases} \\ \tilde{c}_{2k+1} \quad \Rightarrow & \begin{cases} \tilde{c}_{4k+1} \quad \Rightarrow & \begin{cases} \tilde{c}_{8k+1} \\ \tilde{c}_{8k+5} \end{cases} \\ \tilde{c}_{4k+3} \quad \Rightarrow & \begin{cases} \tilde{c}_{8k+3} \\ \tilde{c}_{8k+7} \end{cases} \end{cases} \end{cases} \tag{19.233}$$

$(k = 0, 1, \ldots, 7) \quad (k = 0, 1, 2, 3) \quad (k = 0, 1) \quad (k = 0).$

If the coefficients $\tilde{c}_k$ are substituted into **Scheme 1** and one considers the binary forms of the indices before step 1 and after step 3, then it is easy to recognize that the order of the required coefficients can be obtained by simply *reversing the order of the bits* of the binary form of their indices. This is shown in **Scheme 3**.

Scheme 3:

	Index	Step 1	Step 2	Step 3	Index
$\tilde{c}_0$	000	$\tilde{c}_0$	$\tilde{c}_0$	$\tilde{c}_0$	000
$\tilde{c}_1$	00L	$\tilde{c}_2$	$\tilde{c}_4$	$\tilde{c}_4$	L00
$\tilde{c}_2$	0L0	$\tilde{c}_4$	$\tilde{c}_2$	$\tilde{c}_2$	0L0
$\tilde{c}_3$	0LL	$\tilde{c}_6$	$\tilde{c}_6$	$\tilde{c}_6$	LL0
$\tilde{c}_4$	L00	$\tilde{c}_1$	$\tilde{c}_1$	$\tilde{c}_1$	00L
$\tilde{c}_5$	L0L	$\tilde{c}_3$	$\tilde{c}_5$	$\tilde{c}_5$	L0L
$\tilde{c}_6$	LL0	$\tilde{c}_5$	$\tilde{c}_3$	$\tilde{c}_3$	0LL
$\tilde{c}_7$	LLL	$\tilde{c}_7$	$\tilde{c}_7$	$\tilde{c}_7$	LLL

■ In the case of the function $f(x) = \begin{cases} 2\pi^2 \text{ for } x = 0, \\ x^2 \text{ for } 0 < x < 2\pi, \end{cases}$ with period 2π, the FFT is used for the discrete Fourier transformation. $N = 8$ is chosen. With $x_\nu = \frac{2\pi}{8}$, $f_\nu = \frac{1}{8}f(x_\nu)$ $(\nu = 0, 1, 2, \ldots, 7)$, $\omega_8 = e^{-\frac{2\pi\mathrm{i}}{8}} = 0.707107(1 - \mathrm{i})$, $\omega_8^2 = -\mathrm{i}$, $\omega_8^3 = -0.707107(1 + \mathrm{i})$ **Scheme 4** is got:

Scheme 4:	Step 1	Step 2	Step 3
$f_0 = 2.467401$	$y_0 = 3.701102$	$y_0 = 6.785353$	$y_0 = 13.262281 = \tilde{c}_0$
$f_1 = 0.077106$	$y_1 = 2.004763$	$y_1 = 6.476928$	$y_1 = 0.308425 = \tilde{c}_4$
$f_2 = 0.308425$	$y_2 = 3.084251$	$y_2 = 0.616851$	$y_2 = 0.616851 + 2.467402\,\mathrm{i} = \tilde{c}_2$
$f_3 = 0.693957$	$y_3 = 4.472165$	$y_3 = 2.467402\,\mathrm{i}$	$y_3 = 0.616851 - 2.467402\,\mathrm{i} = \tilde{c}_6$
$f_4 = 1.233701$	$y_4 = 1.233700$	$y_4 = 1.233700 + 2.467401\,\mathrm{i}$	$y_4 = 2.106058 + 5.956833\,\mathrm{i} = \tilde{c}_1$
$f_5 = 1.927657$	$y_5 = -1.308537(1-\mathrm{i})$	$y_5 = 0.872358 + 3.489432\,\mathrm{i}$	$y_5 = 0.361342 - 1.022031\,\mathrm{i} = \tilde{c}_5$
$f_6 = 2.775826$	$y_6 = 2.467401\,\mathrm{i}$	$y_6 = 1.233700 - 2.467401\,\mathrm{i}$	$y_6 = 0.361342 + 1.022031\,\mathrm{i} = \tilde{c}_3$
$f_7 = 3.778208$	$y_7 = 2.180895(1+\mathrm{i})$	$y_7 = -0.872358 + 3.489432\,\mathrm{i}$	$y_7 = 2.106058 - 5.956833\,\mathrm{i} = \tilde{c}_7$

From the third (last) reduction step the required real Fourier coefficients are obtained according to (19.223). (See the right-hand side.)
In this example, the general property

$a_0 = 26.524\,562$	
$a_1 = 4.212\,116$	$b_1 = -11.913\,666$
$a_2 = 1.233\,702$	$b_2 = -\ 4.934\,804$
$a_3 = 0.722\,684$	$b_3 = -\ 2.044\,062$
$a_4 = 0.616\,850$	$b_4 = 0$

$$\tilde{c}_{N-k} = \bar{\tilde{c}}_k \qquad (19.234)$$

of the discrete complex Fourier coefficients can be observed. For $k = 1, 2, 3$, it can be observed that $\tilde{c}_7 = \bar{\tilde{c}}_1$, $\tilde{c}_6 = \bar{\tilde{c}}_2$, $\tilde{c}_5 = \bar{\tilde{c}}_3$.

19.7 Representation of Curves and Surfaces with Splines

19.7.1 Cubic Splines

Since interpolation and approximation polynomials of higher degree usually have unwanted oscillations, it is useful to divide the approximation interval into subintervals by the so-called *nodes* and to consider a relatively simple approximation function on every subinterval. In practice, cubic polynomials are mostly used. A smooth transition is required at the nodes of this piecewise approximation.

19.7.1.1 Interpolation Splines

1. Definition of the Cubic Interpolation Splines, Properties

Suppose there are given N interpolation points (x_i, f_i) $(i = 1, 2, \ldots, N;\ x_1 < x_2 < \ldots x_N)$. The *cubic interpolation spline* $S(x)$ is determined uniquely by the following properties:

1. $S(x)$ satisfies the interpolation conditions $S(x_i) = f_i$ $(i = 1, 2, \ldots, N)$.

2. $S(x)$ is a polynomial of degree ≤ 3 in any subinterval $[x_i, x_{i+1}]$ $(i = 1, 2, \ldots, N-1)$.

3. $S(x)$ is twice continuously differentiable in the entire approximation interval $[x_1, x_N]$.

4. $S(x)$ satisfies the special boundary conditions:

- **a)** $S''(x_1) = S''(x_N) = 0$ (we call them *natural splines*) or
- **b)** $S'(x_1) = f_1{}'$, $S'(x_N) = f_N{}'$ ($f_1{}'$ and $f_N{}'$ are given values) or
- **c)** $S(x_1) = S(x_N)$, in the case of $f_1 = f_N$, $S'(x_1) = S'(x_N)$ and $S''(x_1) = S''(x_N)$ (they are called *periodic splines*).

It follows from these properties that for all twice continuously differentiable functions $g(x)$ satisfying the interpolation conditions $g(x_i) = f_i$ $(i = 1, 2, \ldots, N)$

$$\int_{x_1}^{x_N} [S''(x)]^2\, dx \le \int_{x_1}^{x_N} [g''(x)]^2\, dx \tag{19.235}$$

is valid (*Holladay's Theorem*). Based on (19.235) one can say that $S(x)$ has *minimal total curvature*, since for the curvature κ of a given curve, in a first approximation, $\kappa \approx S''$ (see 3.6.1.2, **4.**, p. 246). It can be shown that if a thin elastic ruler (its name is spline) is led through the points (x_i, f_i) $(i = 1, 2, \ldots, N)$, its bending line follows the cubic spline $S(x)$.

2. Determination of the Spline Coefficients

The cubic interpolation spline $S(x)$ for $x \in [x_i, x_{i+1}]$ has the form:

$$S(x) = S_i(x) = a_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3 \quad (i = 1, 2, \ldots, N - 1). \tag{19.236}$$

The length of the subinterval is denoted by $h_i = x_{i+1} - x_i$. The coefficients of the natural spline can be determined in the following way:

1. From the interpolation conditions we get

$$a_i = f_i \quad (i = 1, 2, \ldots, N - 1). \tag{19.237}$$

It is reasonable to introduce the additional coefficient $a_N = f_N$, which does not occur in the polynomials.

2. The continuity of $S''(x)$ at the interior nodes requires that

$$d_{i-1} = \frac{c_i - c_{i-1}}{3h_{i-1}} \quad (i = 2, 3, \ldots, N - 1). \tag{19.238}$$

The natural conditions result in $c_1 = 0$, and (19.238) still holds for $i = N$, if $c_N = 0$ is introduced.

3. The continuity of $S(x)$ at the interior nodes results in the relation

$$b_{i-1} = \frac{a_i - a_{i-1}}{h_{i-1}} - \frac{2c_{i-1} + c_i}{3} h_{i-1} \quad (i = 2, 3, \ldots, N). \tag{19.239}$$

4. The continuity of $S'(x)$ at the interior nodes requires that

$$c_{i-1}h_{i-1} + 2(h_{i-1} + h_i)c_i + c_{i+1}h_i = 3\left(\frac{a_{i+1} - a_i}{h_i} - \frac{a_i - a_{i-1}}{h_{i-1}}\right) \quad (i = 2, 3, \ldots, N - 1). \tag{19.240}$$

Because of (19.237), the right-hand side of the linear equation system (19.240) to determine the coefficients c_i $(i = 2, 3, \ldots, N - 1;\ c_1 = c_N = 0)$ is known. The left hand-side has the following form:

$$\begin{pmatrix} 2(h_1 + h_2) & h_2 & & & \\ h_2 & 2(h_2 + h_3) & h_3 & \mathbf{O} & \\ & h_3 & 2(h_3 + h_4) & h_4 & \\ & & \ddots & \ddots & \ddots \\ & \mathbf{O} & & & h_{N-2} \\ & & & h_{N-2} & 2(h_{N-2} + h_{N-1}) \end{pmatrix} \begin{pmatrix} c_2 \\ c_3 \\ c_4 \\ \vdots \\ \\ c_{N-1} \end{pmatrix}. \tag{19.241}$$

The coefficient matrix is *tridiagonal*, so the system of equations (19.240) can be solved numerically very easily by an LR decomposition (see 19.2.1.1, **2.**, p. 956). Then all other coefficients in (19.239) and (19.238) can be determined with these values c_i.

19.7.1.2 Smoothing Splines

The given function values f_i are usually measured values in practical applications so they have some error. In this case, the interpolation requirement is not reasonable. This is the reason why *cubic smoothing splines* are introduced. This spline is obtained if in the cubic interpolation splines the interpolation

requirements are replaced by

$$\sum_{i=1}^{N}\left[\frac{f_i - S(x_i)}{\sigma_i}\right]^2 + \lambda\int_{x_1}^{x_N}[S''(x)]^2\,dx = \min!. \tag{19.242}$$

The requirements of continuity of S, S' and S'' are kept, so the determination of the coefficients is a constrained optimization problem with conditions given in equation form. The solution can be obtained by using a Lagrange function (see 6.2.5.6, p. 456). For details see [19.26].

In (19.242) λ ($\lambda \geq 0$) represents a *smoothing parameter*, which must be given previously. For $\lambda = 0$ the result is the cubic interpolation spline, as a special case. For "large" λ the result is a smooth approximation curve, but it returns the measured values inaccurately, and for $\lambda = \infty$ the result is the approximating regression line as another special case. A suitable choice of λ can be made, e.g., by computer-screen dialog. The parameter σ_i ($\sigma_i > 0$) in (19.242) represents the *standard deviation* (see 16.4.1.3, **2.**, p. 851) of the measurement errors, of the values f_i ($i = 1, 2, \ldots, N$).

Until now, the abscissae of the interpolation points and the measurement points were the same as the nodes of the spline function. For large N this method results in a spline containing a large number of cubic functions (19.236). A possible solution is to choose the number and the position of the nodes freely, because in many practical applications only a few spline segments are satisfactory. It is reasonable also from a numerical viewpoint to replace (19.236) by a spline of the form

$$S(x) = \sum_{i=1}^{r+2} a_i N_{i,4}(x). \tag{19.243}$$

Here r is the number of freely chosen nodes, and the functions $N_{i,4}(x)$ are the so-called *normalized B-splines* (*basis splines*) of order 4, i.e., polynomials of degree three, with respect to the i-th node. For details see [19.5].

19.7.2 Bicubic Splines

19.7.2.1 Use of Bicubic Splines

Bicubic splines are used for the following problem: A rectangle R of the x, y plane, given by $a \leq x \leq b$, $c \leq y \leq d$, is decomposed by the *grid points* (x_i, y_j) ($i = 0, 1, \ldots, n;\ j = 0, 1, \ldots, m$) with

$$a = x_0 < x_1 < \cdots < x_n = b, \quad c = y_0 < y_1 < \cdots < y_m = d \tag{19.244}$$

into *subdomains* R_{ij}, where the subdomain R_{ij} contains the points (x, y) with $x_i \leq x \leq x_{i+1}$, $y_j \leq y \leq y_{j+1}$ ($i = 0, 1, \ldots, n-1;\ j = 0, 1, \ldots, m-1$). The values of the function $f(x, y)$ are given at the grid points

$$f(x_i, y_j) = f_{ij} \quad (i = 0, 1, \ldots, n;\ j = 0, 1, \ldots, m). \tag{19.245}$$

A possible simple, smooth surface over R is required which approximates the points (19.245).

19.7.2.2 Bicubic Interpolation Splines

1. Properties

The bicubic interpolation spline $S(x, y)$ is defined uniquely by the following properties:

1. $S(x, y)$ satisfies the interpolation conditions

$$S(x_i, y_j) = f_{ij} \quad (i = 0, 1, \ldots, n;\ j = 0, 1, \ldots, m). \tag{19.246}$$

2. $S(x, y)$ is identical to a bicubic polynomial on every R_{ij} of the rectangle R, that is,

$$S(x, y) = S_{ij}(x, y) = \sum_{k=0}^{3}\sum_{l=0}^{3} a_{ijkl}(x - x_i)^k (y - y_j)^l \tag{19.247}$$

on R_{ij}. So, $S_{ij}(x, y)$ is determined by 16 coefficients, and for the determination of $S(x, y)$ $16 \cdot m \cdot n$ coefficients are needed.

3. The derivatives

$$\frac{\partial S}{\partial x}, \quad \frac{\partial S}{\partial y}, \quad \frac{\partial^2 S}{\partial x \partial y} \tag{19.248}$$

are continuous on R. So, a certain smoothness is ensured for the entire surface.

4. $S(x, y)$ satisfies the special boundary conditions:

$$\begin{aligned} \frac{\partial S}{\partial x}(x_i, y_j) &= p_{ij} \quad \text{for} \quad i = 0, n;\ j = 0, 1, \ldots, m, \\ \frac{\partial S}{\partial y}(x_i, y_j) &= q_{ij} \quad \text{for} \quad i = 0, 1, \ldots, n;\ j = 0, m, \\ \frac{\partial^2 S}{\partial x \partial y}(x_i, y_j) &= r_{ij} \quad \text{for} \quad i = 0, n;\ j = 0, m. \end{aligned} \tag{19.249}$$

Here p_{ij}, q_{ij} and r_{ij} are previously given values.

The results of one-dimensional cubic spline interpolation can be used for the determination of the coefficients a_{ijkl}.

1. There is a very large number $(2n + m + 3)$ of linear systems of equations but only with tridiagonal coefficient matrices.

2. The linear systems of equations differ from each other only on their right-hand sides.

In general, it can be said that bicubic interpolation splines are useful with respect to computation cost and accuracy, and so they are appropriate procedures for practical applications. For practical methods of computing the coefficients see the literature.

2. Tensor Product Approach

The bicubic spline approach (19.247) is an example of the so-called *tensor product* approach having the form

$$S(x, y) = \sum_{i=0}^{n} \sum_{j=0}^{m} a_{ij} g_i(x) h_j(y) \tag{19.250}$$

and which is especially suitable for approximations over a rectangular grid. The functions $g_i(x)$ $(i = 0, 1, \ldots, n)$ and $h_j(y)$ $(j = 0, 1, \ldots, m)$ form two linearly independent function systems. The tensor product approach has the big advantage, from numerical viewpoint, that, e.g., the solution of a two-dimensional interpolation problem (19.246) can be reduced to a one-dimensional one. Furthermore, the two-dimensional interpolation problem (19.246) is uniquely solvable with the approach (19.250) if

1. the one-dimensional interpolation problem with functions $g_i(x)$ with respect to the interpolation nodes $x_0, x_1, \ldots, x_n$ and

2. the one-dimensional interpolation problem with functions $h_j(y)$ with respect to the interpolation nodes $y_0, y_1, \ldots, y_m$

are uniquely solvable.

An important tensor product approach is that with the cubic B-splines:

$$S(x, y) = \sum_{i=1}^{r+2} \sum_{j=1}^{p+2} a_{ij} N_{i,4}(x) N_{j,4}(y). \tag{19.251}$$

Here, the functions $N_{i,4}(x)$ and $N_{j,4}(y)$ are normalized B-splines of order four. Here r denotes the number of nodes with respect to x, p denotes the number of nodes with respect to y. The nodes can be chosen freely but their positions must satisfy certain conditions for the solvability of the interpolation problem.

The B-spline approach results in a system of equations with a band structured coefficient matrix, which is a numerically useful structure.

For solutions of different interpolation problems using bicubic B-splines see the literature.

19.7.2.3 Bicubic Smoothing Splines

The one-dimensional cubic approximation spline is mainly characterized by the optimality condition (19.242). For the two-dimensional case there could be determined a whole sequence of corresponding optimality conditions, however only a few special cases make the existence of a unique solution possible. For appropriate optimality conditions and algorithms for solution of the approximation problem with bicubic B-splines see the literature.

19.7.3 Bernstein–Bézier Representation of Curves and Surfaces

1. Bernstein Basis Polynomials

The Bernstein–Bézier representation (briefly B–B representation) of curves and surfaces applies the *Bernstein polynomials*

$$B_{i,n}(t) = \binom{n}{i} t^i (1-t)^{n-i} \quad (i = 0, 1, \ldots, n) \tag{19.252}$$

and uses the following fundamental properties:

1. $0 \le B_{i,n}(t) \le 1$ for $0 \le t \le 1$, (19.253)

2. $\sum_{i=0}^{n} B_{i,n}(t) = 1.$ (19.254)

Formula (19.254) follows directly from the binomial theorem (see 1.1.6.4, p. 12).

■ **A:** $B_{01}(t) = 1 - t\,,\ B_{1,1}(t) = t$ **(Fig. 19.12)**.

■ **B:** $B_{03}(t) = (1-t)^3\,,\ B_{1,3}(t) = 3t(1-t)^2\,,\ B_{2,3}(t) = 3t^2(1-t)\ B_{3,3}(t) = t^3$ **(Fig. 19.13)**.

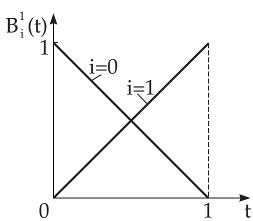


Figure 19.12

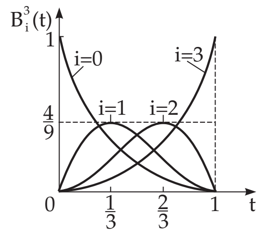


Figure 19.13

2. Vector Representation

Now, a space curve, whose parametric representation is $x = x(t)$, $y = y(t)$, $z = z(t)$, will be denoted in vector form by

$$\vec{\mathbf{r}} = \vec{\mathbf{r}}(t) = x(t)\,\vec{\mathbf{e}}_x + y(t)\,\vec{\mathbf{e}}_y + z(t)\,\vec{\mathbf{e}}_z. \tag{19.255}$$

Here t is the parameter of the curve. The corresponding representation of a surface is

$$\vec{\mathbf{r}} = \vec{\mathbf{r}}(u,v) = x(u,v)\,\vec{\mathbf{e}}_x + y(u,v)\,\vec{\mathbf{e}}_y + z(u,v)\,\vec{\mathbf{e}}_z. \tag{19.256}$$

Here, u and v are the surface parameters.

19.7.3.1 Principle of the B–B Curve Representation

Suppose there are given $n+1$ vertices P_i $(i = 0, 1, \ldots, n)$ of a three-dimensional polygon with the position vectors $\vec{\mathbf{P}}_i$. Introducing the vector-valued function

$$\vec{\mathbf{r}}(t) = \sum_{i=0}^{n} B_{i,n}(t)\vec{\mathbf{P}}_i \tag{19.257}$$

a space curve is assigned to these points, which is called the B–B curve. Because of (19.254) formula (19.257) can be considered as a "variable convex combination" of the given points. The three-dimensional curve (19.257) has the following important properties:

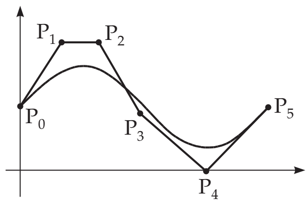


Figure 19.14

1. The points P_0 and P_n are interpolated.

2. Vectors $\overrightarrow{P_0P_1}$ and $\overrightarrow{P_{n-1}P_n}$ are tangents to $\vec{\mathbf{r}}(t)$ at points P_0 and P_n.

The relation between a polygon and a B–B curve is shown in **Fig. 19.14**.

The B–B representation is considered as a design of the curve, since it is easy to influence the shape of the curve by changing the polygon vertices.

Often normalized B-splines are used instead of Bernstein polynomials.

The corresponding space curves are called the B-spline curves. Their shape corresponds basically to the B–B curves with the following advantages:

1. The polygon is better approximated.

2. The B-spline curve changes only locally if the polygon vertices are changed.

3. In addition to the local changes of the shape of the curve the differentiability can also be influenced. So, it is possible to produce break points and line segments for example.

19.7.3.2 B–B Surface Representation

Suppose there are given the points P_{ij} $(i = 0, 1, \ldots, n;\ j = 0, 1, \ldots, m)$ with the position vectors $\vec{\mathbf{P}}_{ij}$, which can be considered as the nodes of a grid along the parameter curves of a surface. Analogously to the B–B curves (19.257), a surface is assigned to the grid points by

$$\vec{\mathbf{r}}(u, v) = \sum_{i=0}^{n} \sum_{j=0}^{m} B_{i,n}(u) B_{j,m}(v) \vec{\mathbf{P}}_{ij}. \tag{19.258}$$

Representation (19.258) is useful for surface design, since by changing the grid points the surface can be changed. Anyway, the influence of every grid point is global, so one should change from the Bernstein polynomials to the B-splines in (19.258).

19.8 Using the Computer

19.8.1 Internal Symbol Representation

Computers are machines that work with symbols. The interpretation and processing of these symbols is determined and controlled by the software. The external symbols, letters, cyphers and special symbols are internally represented in binary code by a form of bit sequence. A *bit* (binary digit) is the smallest representable information unit with values 0 and 1. Eight bits form the next unit, the *byte*. In a byte one can distinguish between 2^8 bit combinations, so 256 symbols can be assigned to them. Such an assignment is called a *code*. There are different codes; one of the most widespread is **ASCII** (**A**merican **S**tandard **C**ode for **I**nformation **I**nterchange).

19.8.1.1 Number Systems

1. Law of Representation

Numbers are represented in computers in a sequence of consecutive bytes. The basis for the internal representation is the binary system, which belongs to the polyadic systems, similarly to the decimal system.

The law of representation for a polyadic number system is

$$a = \sum_{i=-m}^{n} z_i B^i \quad (m > 0,\ n \geq 0;\ m, n \text{ integer}) \tag{19.259}$$

with B as basis and z_i $(0 \leq z_i < B)$ as a digit of the number system. The positions $i \geq 0$ form the integers, those with $i < 0$ the fractional part of the number.

■ The *decimal number representation*, i.e., $B = 10$, of the decimal number 139.8125 has the form $139.8125 = 1 \cdot 10^2 + 3 \cdot 10^1 + 9 \cdot 10^0 + 8 \cdot 10^{-1} + 1 \cdot 10^{-2} + 2 \cdot 10^{-3} + 5 \cdot 10^{-4}$.

The *number systems* occurring most often in computers are shown in **Table 19.3**.

Table 19.3 Number systems

Number system	Basis	Corresponding digits
Binary system	2	$0, 1$
Octal system	8	$0, 1, 2, 3, 4, 5, 6, 7$
Hexadecimal system	16	$0, 1, 2, 3, 4, 5, 6, 7, 8, 9$, A,B,C,D,E,F (The letters A–F are for the values 10–15.)
Decimal system	10	$0, 1, 2, 3, 4, 5, 6, 7, 8, 9$

2. Conversion

The transition from one number system to another is called *conversion*. If different number systems are used in the same time, in order to avoid confusion the basis is denoted as an index.

■ The decimal number 139.8125 is in different systems: $139.8125_{10} = 10001011.1101_2 = 213.64_8 = 8\text{B.D}_{16}$.

1. Conversion of Binary Numbers into Octal or Hexadecimal Numbers The conversion of binary numbers into octal or hexadecimal numbers is simple. Groups of three or four bits are formed starting at the binary point to the left and to the right, and their values are determined. These values are the digits of the octal or hexadecimal systems.

2. Conversion of Decimal Numbers into Binary, Octal or Hexadecimal Numbers For the conversion of a decimal numbers into another system, the following rules are applied for the integer and for the fractional part separately:

a) Integer Part: If G is an integer in the decimal system, then for the number system with basis B the law of formation (19.259) is:

$$G = \sum_{i=0}^{n} z_i B^i \quad (n \geq 0). \tag{19.260}$$

If G is divided by B, then an integer part (the sum) is obtained and a residue:

$$\frac{G}{B} = \sum_{i=1}^{n} z_i B^{i-1} + \frac{z_0}{B}. \tag{19.261}$$

Here, z_0 can have the values $0, 1, \ldots, B-1$, and it is the lowest valued digit of the required number. If this method is repeated for the quotients, further digits can be got.

b) Fractional Part: If g is a proper fraction, then the method to convert it into the number system with basis B is

$$gB = z_{-1} + \sum_{i=2}^{m} z_{-i} B^{-i+1}, \tag{19.262}$$

i.e., the next digit is obtained as the integer part of the product gB. The values z_{-2}, z_{-3}, ... can be obtained in the same way.

■ **A:** Conversion of the decimal number 139 into a binary number.

$139 : 2 = 69$	residue	1	$(1 = z_0)$
$69 : 2 = 34$	residue	1	$(1 = z_1)$
$34 : 2 = 17$	residue	0	$(0 = z_2)$
$17 : 2 = 8$	residue	1	$\vdots$
$8 : 2 = 4$	residue	0	$\vdots$
$4 : 2 = 2$	residue	0	$\vdots$
$2 : 2 = 1$	residue	0	$\vdots$
$1 : 2 = 0$	residue	1	$(1 = z_7)$

$139_{10} = 10001011_2$

■ **B:** Conversion of a decimal fraction 0.8125 into a binary fraction.

$0.8125 \cdot 2 = 1.625$	$(1 = z_{-1})$
$0.625 \cdot 2 = 1.25$	$(1 = z_{-2})$
$0.25 \cdot 2 = 0.5$	$(0 = z_{-3})$
$0.5 \cdot 2 = 1.0$	$(1 = z_{-4})$
$0.0 \cdot 2 = 0.0$	

$0.8125_{10} = 0.1101_2$

3. Conversion of Binary, Octal, and Hexadecimal Numbers into a Decimal Number The algorithm for the conversion of a value from the binary, octal, or hexadecimal system into the decimal system is the following, where the decimal point is after z_0:

$$a = \sum_{i=-m}^{n} z_i B^i \quad (m > 0,\ n \geq 0, \text{integer}). \tag{19.263}$$

The calculation is convenient with the Horner rule (see 19.1.2.1, p. 952).

■ $LLLOL = 1 \cdot 2^4 + 1 \cdot 2^3 + 1 \cdot 2^2 + 0 \cdot 2^1 + 1 \cdot 2^0 = 29.$

The corresponding Horner scheme is shown on the right.

	1	1	1	0	1
2		2	6	14	28
	1	3	7	14	29

19.8.1.2 Internal Number Representation INR

Binary numbers are represented in computers in one or more bytes. Two types of form of representation are distinguished, the *fixed-point numbers* and the *floating-point numbers*. In the first case, the decimal point is at a fixed place, in the second case it is "floating" with the change of the exponent.

1. Fixed-Point Numbers
The range for fixed-point numbers with the given parameters is

$$0 \leq |\, a \,| \leq 2^t - 1. \tag{19.264}$$

Fixed-point numbers can be represented in the form of **Fig. 19.15**.

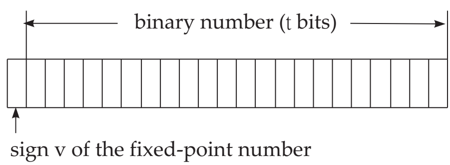


Figure 19.15

2. Floating-Point Numbers
Basically, two different forms are in use for the representation of floating-point numbers, where the internal implementation can vary in detail.

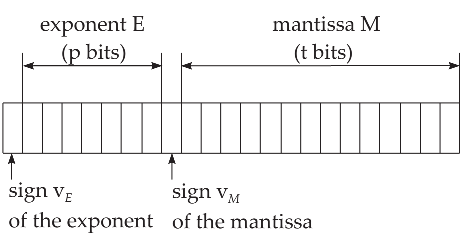


Figure 19.16

1. Normalized Semilogarithmic Form
In the first form, the signs of the exponent E and the mantissa M of the number a are stored separately

$$a = \pm M B^{\pm E}. \tag{19.265a}$$

Here the exponent E is chosen so that for the mantissa

$$1/B \leq M < 1 \tag{19.265b}$$

holds. It is called the *normalized semilogarithmic form* **(Fig. 19.16)**.

The range of the absolute value of the floating-point numbers with the given parameters is:

$$2^{-2^p} \le |\, a \,| \le \left(1 - 2^{-t}\right) \cdot 2^{(2^p-1)}. \tag{19.266}$$

2. IEEE Standard The second (nowadays used) form of floating-point numbers corresponds to the **IEEE** (**I**nstitute of **E**lectrical and **E**lectronics **E**ngineers) *standard* accepted in 1985. It deals with the requirements of computer arithmetic, roundoff behavior, arithmetical operators, conversion of numbers, comparison operators and handling of exceptional cases such as over- and underflow.

The floating-point number representations are shown in **Fig. 19.17**.

The characteristic C comes from the exponent E by addition of a suitable constant K. This is chosen so that only positive numbers occur in the characteristic. The representable number is

$$a = (-1)^v \cdot 2^E \cdot 1.b_1b_2\ldots b_{t-1} \quad \text{with } E = C - K. \tag{19.267}$$

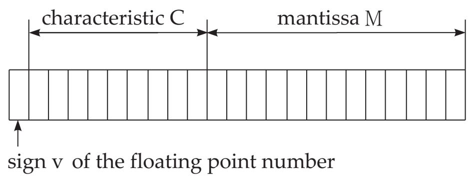


Figure 19.17

Here: $C_{min} = 1$, $C_{max} = 254$, since $C = 0$ and $C = 255$ are reserved.
The standard gives two basic forms of representation (single-precision and double-precision floating-point numbers), but other representations are also possible. **Table 19.4** contains the parameters for the basic forms.

Table 19.4 Parameters for the basic forms

Parameter	**Single precision**	**Double precision**
Word length in bits	32	64
Maximal exponent E_{max}	+127	+1023
Minimal exponent E_{min}	−126	−1022
Constant K	+127	+1023
Number of bits in exponent	8	11
Number of bits in mantissa	24	53

19.8.2 Numerical Problems in Calculations with Computers

19.8.2.1 Introduction, Error Types

The general properties of calculations with a computer are basically the same as those of calculations done by hand, however some of them need special attention, because the accuracy comes from the representation of the numbers, and from the missing judgement with respect to the errors of the computer. Furthermore, computers perform many more calculation steps than human can do manually.

So, there is the problem of how to influence and control the errors, e.g., by choosing the most appropriate numerical method among the mathematically equivalent methods.

In further discussions, the following notation is used, where x denotes the exact value of a quantity, which is mostly unknown, and $\tilde{x}$ is an approximation value of x:

Absolute error: $|\Delta x| = |x - \tilde{x}|$. (19.268) Relative error: $\left|\dfrac{\Delta x}{x}\right| = \left|\dfrac{x-\tilde{x}}{x}\right|$. (19.269)

The notations

$$\epsilon(x) = x - \tilde{x} \quad \text{and} \quad \epsilon_{rel}(x) = \frac{x-\tilde{x}}{x} \tag{19.270}$$

are also often used.

19.8.2.2 Normalized Decimal Numbers and Round-Off

1. Normalized Decimal Numbers

Every real number $x \neq 0$ can be expressed as a decimal number in the form

$$x = \pm 0.b_1 b_2 \ldots \cdot 10^E \quad (b_1 \neq 0). \tag{19.271}$$

Here $0, b_1 b_2 \ldots$ is called the *mantissa* formed with the digits $b_i \in \{0, 1, 2, \ldots, 9\}$. The number E is an integer, the so-called exponent with respect to the base 10. Since $b_1 \neq 0$, (19.271) is called a *normalized decimal number*.

Since only finitely many digits can be handled by a real computer, one has to restrict himself to a fixed number t of mantissa digits and to a fixed range of the exponent E. So, from the number x given in (19.271) the number

$$\tilde{x} = \begin{cases} \pm 0.b_1 b_2 \cdots b_t \cdot 10^E & \text{for} \quad b_{t+1} \leq 5 \ \text{(round-down)}, \\ \pm(0.b_1 b_2 \cdots b_t + 10^{-t}) 10^E & \text{for} \quad b_{t+1} > 5 \ \text{(round-up)}, \end{cases} \tag{19.272}$$

is obtained by round-off (as it is usual in practical calculations). The absolute error caused by round-off

$$|\Delta x| = |x - \tilde{x}| \leq 0.5 \cdot 10^{-t} 10^E. \tag{19.273}$$

2. Basic Operations and Numerical Calculations

Every numerical process is a sequence of basic calculation operations. Problems arise especially with the finite number of positions in the floating-point representation. Here a short overview is given. It is supposed that x and y are normalized error-free floating-point numbers with the same sign and with a non-zero value:

$$x = m_1 B^{E_1}, \quad y = m_2 B^{E_2} \quad \text{with} \tag{19.274a}$$

$$m_i = \sum_{k=1}^{t} a_{-k}^{(i)} B^{-k}, \quad a_{-1}^{(i)} \neq 0, \quad \text{and} \tag{19.274b}$$

$$a_{-k}^{(i)} = 0 \text{ or } 1 \text{ or} \ldots \text{ or} \quad B - 1 \ \text{ for} \quad k > 1 \quad (i = 1, 2). \tag{19.274c}$$

1. Addition If $E_1 > E_2$, then the common exponent becomes E_1, since normalization allows us to make only a left-shift. The mantissas are then added.

If $\quad B^{-1} \leq | m_1 + m_2 B^{-(E_1 - E_2)} | < 2 \qquad (19.275a) \qquad$ and $\quad | m_1 + m_2 B^{-(E_1 - E_2)} | \geq 1, \qquad (19.275b)$

then shifting the decimal point by one position to the left results in an increase of the exponent by one.

■ $0.9604 \cdot 10^3 + 0.5873 \cdot 10^2 = 0.9604 \cdot 10^3 + 0.05873 \cdot 10^3 = 1.01913 \cdot 10^3 = 0.1019 \cdot 10^4$.

2. Subtraction The exponents are equalized as in the case of addition, the mantissas are then subtracted. If

$| m_1 - m_2 B^{-(E_1 - E_2)} | < 1 - B^{-t} \ (19.276a) \qquad$ and $\quad | m_1 - m_2 B^{-(E_1 - E_2)} | < B^{-1}, \qquad (19.276b)$

shifting the decimal point to the right by a maximum of t positions results in the corresponding decrease of the exponent.

■ $0.1004 \cdot 10^3 - 0.9988 \cdot 10^2 = 0.1004 \cdot 10^3 - 0.09988 \cdot 10^3 = 0.00052 \cdot 10^3 = 0.5200 \cdot 10^0$. This example shows the critical case of subtractive cancellation. Because of the limited number of positions (here four), zeros are carried in from the right instead of the correct characters.

3. Multiplication The exponents are added and the mantissas are multiplied. If

$$m_1 m_2 < B^{-1}, \tag{19.277}$$

then the decimal point is shifted to the right by one position, and the exponent is decreased by one.

■ $(0.3176 \cdot 10^3) \cdot (0.2504 \cdot 10^5) = 0.07952704 \cdot 10^8 = 0.7953 \cdot 10^7$.

4. Division The exponents are subtracted and the mantissas are divided. If

$$\frac{m_1}{m_2} \geq B^{-1}, \tag{19.278}$$

then the decimal point is shifted to the left by one position, and the exponent is increased by one.

■ $(0.3176 \cdot 10^3)/(0.2504 \cdot 10^5) = 1.2683706\ldots 10^{-2} = 0.1268 \cdot 10^{-1}$.

5. Error of the Result The error of the result in the four basic operations with terms that are supposed to be error-free is a consequence of round-off. For the relative error with number of positions t and the base B, the limit is

$$\frac{B}{2}B^{-t}. \tag{19.279}$$

6. Subtractive cancelation As it was mentioned above, the critical operation is the subtraction of nearly equal floating-point numbers. If it is possible, one should avoid this by changing the order of operations, or by using certain identities.

■ $x = \sqrt{1985} - \sqrt{1984} = 0.4455 \cdot 10^2 - 0.4454 \cdot 10^2 = 0.1 \cdot 10^{-1}$ or $x = \sqrt{1985} - \sqrt{1984} = \dfrac{1985 - 1984}{\sqrt{1985} + \sqrt{1984}} = 0.1122 \cdot 10^{-1}$

19.8.2.3 Accuracy in Numerical Calculations

1. Types of Errors

Numerical methods have errors. There are several types of errors, from which the total error of the final result is accumulated **(Fig. 19.18)**.

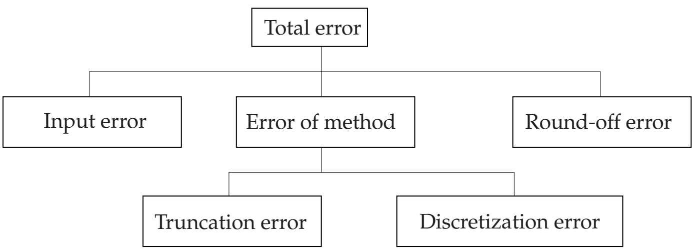


Figure 19.18

2. Input Error

1. Notion of Input Error *Input error* is the error of the result caused by inaccurate input data. Slight inaccuracies of input data are also called *perturbations*. The determination of the error of the input data is called the *direct problem of error calculus*. The *inverse problem* is the following: How large an error the input data may have such that the final input error does not exceed an acceptable tolerance value. The estimation of the input error in rather complex problems is very difficult and is usually hardly possible. In general, for a real-valued function $y = f(\underline{\mathbf{x}})$ with $\underline{\mathbf{x}} = (x_1, x_2, \ldots, x_n)^{\mathrm{T}}$ the absolute value of the input error is

$$\begin{aligned} |\Delta y| &= |f(x_1, x_2, \ldots, x_n) - f(\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n)| \\ &= |\sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(\xi_1, \xi_2, \ldots, \xi_n)(x_i - \tilde{x}_i)| \le \sum_{i=1}^{n} \left(\max_x |\frac{\partial f}{\partial x_i}(\underline{\mathbf{x}})| \right) |\Delta x_i|, \end{aligned} \tag{19.280}$$

if the Taylor formula (see 7.3.3.3, p. 471) is used for $y = f(x) = f(x_1, x_2, \ldots, x_n)$ with a linear residue. $\xi_1, \xi_2, \ldots, \xi_n$ denote the intermediate values, $\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n$ denote the approximating values of $x_1, x_2, \ldots, x_n$. The approximating values are the perturbed input data. Here, also the Gauss error propagation law (see 16.4.2.1, p. 855) is considered.

2. Input Error of Simple Arithmetic Operations The input error is known for simple arithmetical operations. With the notation of (19.268)–(19.270) for the four basic operations:

$$\epsilon(x \pm y) = \epsilon(x) \pm \epsilon(y), \quad (19.281) \qquad \epsilon(xy) = y\epsilon(x) + x\epsilon(y) + \epsilon(x)\epsilon(y), \quad (19.282)$$

$$\epsilon\left(\frac{x}{y}\right) = \frac{1}{y}\epsilon(x) - \frac{x}{y^2}\epsilon(y) + \text{ terms of higher order in } \varepsilon, \quad (19.283)$$

$$\epsilon_{rel}(x \pm y) = \frac{x\epsilon_{rel}(x) \pm y\epsilon_{rel}(y)}{x \pm y}, \quad (19.284) \qquad \epsilon_{rel}(xy) = \epsilon_{rel}(x) + \epsilon_{rel}(y) + \epsilon_{rel}(x)\epsilon_{rel}(y), \quad (19.285)$$

$$\epsilon_{rel}\left(\frac{x}{y}\right) = \epsilon_{rel}(x) - \epsilon_{rel}(y) + \text{ terms of higher order in } \varepsilon. \quad (19.286)$$

The formulas show: Small relative errors of the input data result in small relative errors of the result on multiplication and division. For addition and subtraction, the relative error can be very large if $|x \pm y| \ll |x| + |y|$.

3. Error of the Method

1. Notion of the Error of the Method The *error of the method* comes from the fact that theoretically continuous phenomena are numerically approximated in many different ways as limits. Hence, there are *truncation errors* in limiting processes (as, e.g., in iteration methods) and *discretization errors* in the approximation of continuous phenomena by a finite discrete system (as, e.g., in numerical integration). Errors of methods exist independently of the input and round-off errors; consequently, they can be investigated only in connection with the applied solution methodology.

2. Applying Iteration Methods If an iteration method is used, then both cases may occur: A correct solution or also a false solution of the problem can be obtained. It is also possible that no solution is obtained by an iteration method although there exists one.
To make an iteration method clearer and safer, the following advices should be considered:

a) To avoid "infinite" iterations, count the number of steps and stop the process if this number exceeds a previously given value (i.e., stop without reaching the required accuracy).

b) The location of the intermediate result should be tracked on the screen by a numerical or a graphical representation of the intermediate results.

c) All known properties of the solution should be used such as gradient, monotonicity, etc.

d) The possibilities of scaling the variables and functions should be investigated.

e) Several tests should be performed by varying the step size, truncation conditions, initial values, etc.

4. Round-off Errors

Round-off errors occur because the intermediate results should be rounded. So, they have an essential importance in judging mathematical methods with respect to the required accuracy. They determine together with the errors of input and the error of the method, whether a given numerical method is strongly stable, weakly stable or unstable. Strong *stability*, weak stability, or *instability* occur if the total error, at an increasing number of steps, decreases, has the same order, or increases, respectively. At the instability one distinguishes between the sensitivity with respect to round-off errors and *discretization errors* (numerical instability) and with respect to the error in the initial data at a theoretically exact calculation (natural instability). A calculation process is appropriate if the numerical instability is not greater than the natural instability.

For the local error propagation of round-off errors, i.e., errors at the transition from a calculation step to the next one, the same estimation process can be used as the one applied at the input error.

5. Examples of Numerical Calculations

Some of the problems mentioned above are illustrated by numerical examples.

■ **A: Roots of a Quadratic Equation:**
$ax^2 + bx + c = 0$ with real coefficients a, b, c and $D = b^2 - 4ac \geq 0$ (real roots). Critical situations are the cases **a)** $|\, 4ac \,| \ll b^2$ and **b)** $4ac \approx b^2$. Recommended proceeding:

a) $x_1 = -\dfrac{b+\operatorname{sign}(b)\sqrt{D}}{2a}$, $x_2 = \dfrac{c}{a x_1}$ (Vieta root theorem, see 1.6.3.1, 3., p. 44).

b) The vanishing of D cannot be avoided by a direct method. Subtractive cancellation occurs but the error in $(b + \operatorname{sign}(b\sqrt{D})$ is not too large since $|b| \gg \sqrt{D}$ holds.

■ **B: Volume of a Thin Conical Shell for $h \ll r$**

$V = 4\pi\dfrac{(r+h)^3 - r^3}{3}$ because of $(r+h) \approx r$ there is a case of subtractive cancellation. However in the equation $V = 4\pi\dfrac{3r^2h + 3rh^2 + h^3}{3}$ there is no such problem.

■ **C: Determining the Sum** $S = \sum\limits_{k=1}^{\infty}\dfrac{1}{k^2+1}$ $(S = 1.07667\ldots)$ with an accuracy of three significant digits. Performing the calculations with 8 digits, about 6000 terms should be added. After the identical transformation $\dfrac{1}{k^2+1} = \dfrac{1}{k^2} - \dfrac{1}{k^2(k^2+1)}$

$S = \sum\limits_{k=1}^{\infty}\dfrac{1}{k^2} - \sum\limits_{k=1}^{\infty}\dfrac{1}{k^2(k^2+1)}$ and $S = \dfrac{\pi^2}{6} - \sum\limits_{k=1}^{\infty}\dfrac{1}{k^2(k^2+1)}$ hold. By this transformation only eight terms are considered.

■ **D: Avoiding the $\frac{0}{0}$ Situation** in the function $z = (1 - \sqrt{1+x^2+y^2})\dfrac{x^2-y^2}{x^2+y^2}$ for $x = y = 0$. Multiplying the numerator and the denominator by $(1 + \sqrt{1+x^2+y^2})$ one avoids this situation.

■ **E: Example for an Unstable Recursive Process.** Algorithms with the general form $y_{n+1} = ay_n + by_{n-1}$ $(n = 1, 2, \ldots)$ are stable if the condition $\left|\dfrac{a}{2} \pm \sqrt{\dfrac{a^2}{4} + b}\right| < 1$ is satisfied. The special case $y_{n+1} = -3y_n + 4y_{n-1}$ $(n = 1, 2, \ldots)$ is unstable. If y_0 and y_1 have errors ε and $-\varepsilon$, then for $y_2, y_3, y_4, y_5, y_6, \ldots$ the errors are 7ε, -25ε, 103ε, -409ε, 1639ε, ... The process is instable for the parameters $a = -3$ and $b = 4$.

■ **F: Numerical Integration of a Differential Equation.** The numerical solution for the first-order ordinary differential equation

$$y' = f(x, y) \text{ with } f(x, y) = ay \tag{19.287}$$

and the initial value $y(x_0) = y_0$ will be represented.

a) Natural Instability. Together with the exact solution $y(x)$ for the exact initial values $y(x_0) = y_0$ let $u(x)$ be the solution for a perturbed initial value. Without loss of generality, it may be assumed that the perturbed solution has the form

$$u(x) = y(x) + \varepsilon\,\eta(x), \tag{19.288a}$$

where ε is a parameter with $0 < \varepsilon < 1$ and $\eta(x)$ is the so-called perturbation function. Considering that $u'(x) = f(x, u)$ one gets from the Taylor expansion (see 7.3.3.3, p. 471)

$$u'(x) = f(x, y(x) + \varepsilon\,\eta(x)) = f(x, y) + \varepsilon\,\eta(x)\, f_y(x, y) + \text{terms of higher order} \tag{19.288b}$$

which implies the so-called differential variation equation

$$\eta'(x) = f_y(x, y)\eta(x). \tag{19.288c}$$

The solution of the problem with $f(x, y) = ay$ is

$$\eta(x) = \eta_0\, e^{a(x-x_0)} \text{ with } \eta_0 = \eta(x_0). \tag{19.288d}$$

For $a > 0$ even a small initial perturbation η_0 results in an unboundedly increasing perturbation $\eta(x)$. So, there is a natural instability.

b) Investigation of the Error of the Method in the Trapezoidal Rule. With $a = -1$, the stable differential equation $y'(x) = -y(x)$ has the exact solution

$$y(x) = y_0 e^{-(x-x_0)}, \text{ where } y_0 = y(x_0). \tag{19.289a}$$

The trapezoidal rule is

$$\int_{x_1}^{x_{i+1}} y(x)dx \approx \frac{y_i + y_{i+1}}{2} h \text{ with } h = x_{i+1} - x_i. \tag{19.289b}$$

By using this formula for the given differential equation

$$\tilde{y}_{i+1} = \tilde{y}_i + \int_{x_i}^{x_{i+1}} (-y)dx = \tilde{y}_i - \frac{\tilde{y}_i + \tilde{y}_{i+1}}{2} h \quad \text{or} \quad \tilde{y}_{i+1} = \frac{2-h}{2+h}\tilde{y}_i \quad \text{or}$$

$$\tilde{y}_i = \left(\frac{2-h}{2+h}\right)^i \tilde{y}_0 \tag{19.289c}$$

is valid. With $x_i = x_0 + ih$, i.e., with $i = (x_i - x_0/h$ for $0 \le h < 2$

$$\tilde{y}_i = \left(\frac{2-h}{2+h}\right)^{(x_i-x_0)/h} \tilde{y}_0 = \tilde{y}_0 e^{c(h)(x_i-x_0)} \quad \text{with} \quad c(h) = \frac{\ln\left(\dfrac{2-h}{2+h}\right)}{h} = -1 - \frac{h^2}{12} - \frac{h^4}{80} - \cdots \tag{19.289d}$$

is obtained. If $\tilde{y}_0 = y_0$, then $\tilde{y}_i < y_i$, and so for $h \to 0$, $\tilde{y}_i$ also tends to the exact solution $y_0 e^{-(x_i-x_0)}$.

c) Input Error In **b)** it is supposed that the exact and the approximate initial values coincide. Now, the behavior is investigated when $y_0 \ne \tilde{y}_0$ with $|\ \tilde{y}_0 - y_0\ | < \varepsilon_0$.

$$\text{Since } (\tilde{y}_{i+1} - y_{i+1}) \le \frac{2-h}{2+h}(\tilde{y}_i - y_i) \text{ there is } (\tilde{y}_{i+1} - y_{i+1}) \le \left(\frac{2-h}{2+h}\right)^{i+1} (\tilde{y}_0 - y_0). \tag{19.290a}$$

So, ε_{i+1} is at most of the same order as ε_0, and the method is stable with respect to the initial values. It has to be mentioned that in solving the above differential equation with the Simpson method an artificial instability is introduced. In this case, for $h \to 0$, the general solution is obtained as

$$\tilde{y}_i = C_1 e^{-x_i} + C_2(-1)^i e^{x_i/3}. \tag{19.290b}$$

The problem is that the numerical solution method uses higher-order differences than those to which the order of the differential equation corresponds.

19.8.3 Libraries of Numerical Methods

Over time, *libraries* of functions and procedures have been developed independently of each other for numerical methods in different programming languages. An enormous amount of computer experimentation was considered in their development, so in solutions of practical numerical problems one should use the programs from one of these program libraries. Programs are available for current operating systems like WINDOWS, UNIX and LINUX and mostly for every computation problem type and they keep certain conventions, so it is more or less easy to use them.

The application of methods from program libraries does not relieve the user of the necessity of thinking about the expected results. This is a warning that the user should be informed about the advantages and also about the disadvantages and weaknesses of the mathematical method he/she is going to use.

19.8.3.1 NAG Library

The *NAG library* (**N**umerical **A**lgorithms **G**roup) is a rich collection of numerical methods in the form of functions and subroutines/procedures in the programming languages FORTRAN 77, FORTRAN 90

and C. Here is a contents overview:

1. Complex arithmetic
2. Roots of polynomials
3. Roots of transcendental equations
4. Series
5. Integration
6. Ordinary differential equations
7. Partial differential equations
8. Numeric differentiation
9. Integral equations
10. Interpolation
11. Approximation of curves and surfaces from data
12. Minimum/maximum of a function
13. Matrix operations, inversion
14. Eigenvalues and eigenvectors
15. Determinants
16. Simultaneous linear equations
17. Orthogonalization
18. Linear algebra
19. Simple calculations with statistical data
20. Correlation and regression analysis
21. Random number generators
22. Non-parametric statistics
23. Time series analysis
24. Operations research
25. Special functions
26. Mathematical and computer constants

Furthermore the NAG library contains extensive software concerning statistics and financial mathematics.

19.8.3.2 IMSL Library

The **IMSL library** (**I**nternational **M**athematical and **S**tatistical **L**ibrary) consists of three synchronized parts:

General mathematical methods,
Statistical problems,
Special functions.

The sublibraries contain functions and subroutines in FORTRAN 77, FORTRAN 90 and C. Here is a contents overview:

General Mathematical Methods

1. Linear systems
2. Eigenvalues
3. Interpolation and approximation
4. Integration and differentiation
5. Differential equations
6. Transformations
7. Non-linear equations
8. Optimization
9. Vector and matrix operations
10. Auxiliary functions

Statistical Problems

1. Elementary statistics
2. Regression
3. Correlation
4. Variance analysis
5. Categorization and discrete data analysis
6. Non-parametric statistics
7. Test of goodness of fit and test of randomness
8. Analysis of time series and forecasting
9. Covariance and factor analysis
10. Discriminance analysis
11. Cluster analysis
12. Random sampling
13. Life time distributions and reliability
14. Multidimensional scaling
15. Estimation of reliability function, hazard rate and risk function
16. Line-printer graphics
17. Probability distributions
18. Random number generators
19. Auxiliary algorithms
20. Auxiliary mathematical tools

Special Functions

1. Elementary functions
2. Trigonometric and hyperbolic functions
3. Exponential and related functions
4. Gamma function and relatives
5. Error functions and relatives
6. Bessel functions
7. Kelvin functions
8. Bessel functions with fractional orders
9. Weierstrass elliptic integrals and related functions
10. Different functions

19.8.3.3 Aachen Library

The **Aachen library** is based on the collection of formulas for numerical mathematics of G. Engeln–Müllges (Fachhochschule Aachen) and F. Reutter (Rheinisch–Westfälische Technische Hochschule Aachen). It exists in the programming languages BASIC, QUICKBASIC, FORTRAN 77, FORTRAN 90, C, MODULA 2 and TURBO PASCAL. Here is an overview:

1. Numerical methods to solve non-linear and special algebraic equations
2. Direct and iterative methods to solve systems of linear equations
3. Systems of non-linear equations
4. Eigenvalues and eigenvectors of matrices
5. Linear and non-linear approximation
6. Polynomial and rational interpolation, polynomial splines
7. Numerical differentiation
8. Numerical quadrature
9. Initial value problems of ordinary differential equations
10. Boundary value problems of ordinary differential equations

The programs of the Aachen library are especially suitable for the investigation of individual algorithms of numerical mathematics.

19.8.4 Application of Interactive Program Systems and Computeralgebra Systems

19.8.4.1 Matlab

The commercial program system Matlab Matlab (**Mat**rix **Lab**oratory) is an interactive environment for solving mathematically formulated problems and in the same time it is a high level script language for scientific technical computations. The set up priorities are the problems and algorithms of linear algebra. Matlab unifies the convenient well developed implementations of numerical procedures with advanced graphical representation of the results and data. The computations are processed mostly with double precision floating point numbers according to **IEEE**–*standards* (see **Table19.4**, p. 1004). As further alternatives which are compatible to Matlab are the systems Scilab and Octave with free downloads.

1. Functions Survey

There is a short survey of the procedures and functions available in Matlab:

General Mathematical Functions

1. Trigonometry
2. Exponential functions, Logarithms
3. Special functions
4. Complex arithmetic
5. Coordinate transformations
6. Round-off and fractions
7. Discrete mathematics
8. Mathematical constants

Numerical Linear Algebra

1. Manipulation of fields and matrices
2. Special matrices
3. Matrix analyzes (norms, condition)
4. Systems of linear equations
5. Eigenvalues and singular values
6. Matrix factorization
7. Matrix functions
8. Methods for sparse matrices

Numerical Methods

1. Calculation of statistical data
2. Correlation and regression
3. Discrete Fourier transformation
4. Polynomials and splines
5. One- and more-dimensional interpolation
6. Triangulations and decompositions
7. Determination of the convex closures
8. Numerical integration
9. Ordinary differential equations
10. Partial differential equations
11. Non-linear equations
12. Minimization of functions

In addition there are several program packages of Matlab, the so called toolbooxes, which can be applied in the cases of special mathematical classis of problems. As some examples can be mentioned here the curve fitting, filtering, business mathematics, time series analysis, signal and pattern processing, neural networks, optimization, partial differential equations, splines, statistics and wavelets.

In the following paragraphs the possibilities of Matlab are demonstrated by simple examples. The same problems are partially discussed here as in the paragraphs of numerical applications of Mathematica and Maple.

2. Numerical Linear Algebra

After starting with Matlab the command prompt >> appears in the command window to indicate the readiness to accept commands. If a command is not closed by a semicolon, then the result appears in the command window. The basic command to solve a system of linear equations $\mathbf{A}\,\underline{\mathbf{x}} = \underline{\mathbf{b}}$ (see 19.2.1, p. 955) is the backslash operator \.

■ Given matrix $\mathbf{A} = \begin{pmatrix} 1 & 0 & 3 \\ 2 & 1 & 1 \\ 1 & 2 & 3 \end{pmatrix}$ and vector $\underline{\mathbf{b}} = (-2, 3, 2)^T$. For the input

```
>>    A = [1 0 3; 2 1 1; 1 2 3],   b = [-2; 3; 2],   x = A\b,   norm(A * x - b)
```

the output is

$$\mathtt{A} = \begin{matrix} 1 & 0 & 3 \\ 2 & 1 & 1 \\ 1 & 2 & 3 \end{matrix} \quad \mathtt{b} = \begin{matrix} -2 \\ 3 \\ 2 \end{matrix} \quad \mathtt{x} = \begin{matrix} 1.0000 \\ 2.0000 \\ -1.0000 \end{matrix} \quad \mathtt{ans} = \mathtt{8.8818e-016}$$

As the Euclidean norm of the residual shows the obtained solution x (for which not all the digits are shown) satisfies the system of equations with an accuracy allowed by the mashine floating point representation.

If the matrix **A** is quadratic and nonsingular, then the linear system has a unique solution. By the backslash operator \ ordinary the Gaussian elimination is used with column pivoting, i.e., a triangle decomposition $\mathbf{PA} = \mathbf{LR}$ is obtained (see 19.2.1.1, p. 955).

■ The triangle decomposition of **A** can be realized also with the input

```
>>    [L, R, P] = lu(A)
```

giving the output

$$\mathtt{L} = \begin{pmatrix} 1.0000 & 0 & 0 \\ 0.5000 & 1.0000 & 0 \\ 0.5000 & -0.3333 & 1.0000 \end{pmatrix} \quad \mathtt{R} = \begin{pmatrix} 2.0000 & 1.0000 & 1.0000 \\ 0 & 1.5000 & 2.5000 \\ 0 & 0 & 3.3333 \end{pmatrix} \quad \mathtt{P} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

(Here the matrices are given in brackets in order to avoid confusion).

The backslash operator first tests the properties of the coefficient matrix **A**. If **A** is a permutation of a triangle matrix, then the corresponding echelon form is solved. For a symmetric **A** the Cholesky method is applied (see 19.2.1.2, p. 958).

If the condition number of the coefficient matrix **A** is too high, then numerical problems can occure during the solution. Because of this problem, during the procedure Matlab calculates an estimation of the reciprocal value of the condition number, and gives a warning if it is too small.

■ The Hilbert matrix $\mathbf{H} = (h_{ik})$ of order $n = 13$ can serve as an example, where $h_{ik} = 1/(i + k - 1)$.

```
>> x = hilb(13)\ones(13, 1);
  Warning: Matrix is close to singular or badly scaled.
    Results may be inaccurate. RCOND = 2.409320e-017.
```

In the case of overdetermined linear systems the corresponding linear fitting problem is handled by an orthogonalization procedure, i.e. by orthogonal transformations into a QR-decomposition $\mathbf{A} = \mathbf{QR}$

(see 19.2.1.3, p. 958).

■ >> $\texttt{A} = [1\ 0\ 3; 2\ 1\ 1; 1\ 2\ 3; 1\ 1\ -1];\quad \texttt{b} = [-2; 3; 2; 1];\quad \texttt{x} = \texttt{A}\backslash\texttt{b},\quad \texttt{norm}(\texttt{A} * \texttt{x} - \texttt{b})$

$$\texttt{x} = \begin{matrix} 0.4673 \\ 1.4393 \\ -0.4953 \end{matrix} \qquad \texttt{ans} = 2.0508$$

>> $[\texttt{Q}, \texttt{R}] = \texttt{qr}(\texttt{A})$

$$\texttt{Q} = \begin{pmatrix} -0.3780 & -0.4583 & 0.6466 & 0.4785 \\ -0.7559 & -0.2750 & -0.2321 & -0.5469 \\ -0.3780 & 0.8250 & 0.4145 & -0.0684 \\ -0.3780 & 0.1833 & -0.5969 & 0.6836 \end{pmatrix} \qquad \texttt{R} = \begin{pmatrix} -2.6458 & -1.8898 & -2.6458 \\ 0 & 1.5584 & 0.6417 \\ 0 & 0 & 3.5480 \\ 0 & 0 & 0 \end{pmatrix}$$

The backslash operator also gives meaningful results in the cases of underdetermined and rank deficient linear systems of equations. The details of these cases with the ways how to handle large sparse matrices can be found in the corresponding documentation of Matlab and in the introductions [19.20], [19.29].

3. Numerical Solution of Equations

A polynomial

$$p(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0$$

is in Matlab represented by the row vector $(a_n, a_{n-1}, \ldots, a_1, a_0)$ of the coefficients. Several functions are available to handle polynomials.

■ As an example the polynomial value at 1, the derivative (i.e. the coefficient vector of the derivative polynomial) and the roots are determined for the polynomial $p(x) = x^6 + 3x^2 - 5$.

>> $\texttt{p} = [1\ 0\ 0\ 0\ 3\ 0\ -5];$

>> $\texttt{polyval}(\texttt{p}, 1)\qquad \texttt{ans} = -1$

>> $\texttt{polyder}(\texttt{p})\qquad \texttt{ans} = 6\ 0\ 0\ 0\ 6\ 0$

>> $\texttt{roots}(\texttt{p})\qquad \texttt{ans} = 0.8673 + 1.1529\texttt{i}, 0.8673 - 1.1529\texttt{i}, 1.0743,$

$\qquad\qquad -0.8673 + 1.1529\texttt{i}, -0.8673 - 1.1529\texttt{i}, -1.0743$

The roots are determined as the eigenvalues of the corresponding companion matrix.

The command `fzero` is used to find the approximate solutions of nonlinear scalar equations.

■ Calculation of three solutions of equation $e^{-x^3} - 4x^2 = 0$.

>> $\texttt{fzero}(@(\texttt{x})\texttt{exp}(-\texttt{x}\hat{}3) - 4 * \texttt{x}\hat{}2, 1)\qquad \texttt{ans} = 0.4741$

>> $\texttt{fzero}(@(\texttt{x})\texttt{exp}(-\texttt{x}\hat{}3) - 4 * \texttt{x}\hat{}2, 0)\qquad \texttt{ans} = -0.5413$

>> $\texttt{fzero}(@(\texttt{x})\texttt{exp}(-\texttt{x}\hat{}3) - 4 * \texttt{x}\hat{}2, -1)\qquad \texttt{ans} = -1.2085$

The input of the equation for the command `fzero` is made as an unnamed function. As it is obvious, it depends on the given initial value in the second argument, which solution is approximated. A combination of the bisection method with regula falsi (see 19.1.1.3, p. 951) is used for the iteration process.

4. Interpolation

Function fitting based on a given data set can be done either by interpolation (see 19.6.1, p. 982 or 19.7.1.1, p. 996) or by best approximating functions (see 19.6.2.2, p. 985). In Matlab the command `plot` is the most simple way to represent the data set graphically. The menu in the selfopening graphical window contains tools for editing the figure (linetypes, symbols, titles and legends), for exporting and printing *Basic Fitting* under *Tools*.

The Basic Fitting is a subroutine of Tools by which a variety of interpolation methods and best approximating polynomials of different degrees are offered. It is realized by the functions `interp1` and `polyfit`.

■ By the input

>> `plot([1.70045, 1.2523, 0.638803, 0.423479, 0.249091, 0.160321, 0.0883432, 0.0570776, 0.0302744, 0.0212794]);`

the data values are located at the data positions $1, 2, \ldots, 10$ and are graphically represented. **Fig.19.19a** shows the data set, the corresponding cubic interpolation spline and the best approximating polynomial of degree four.

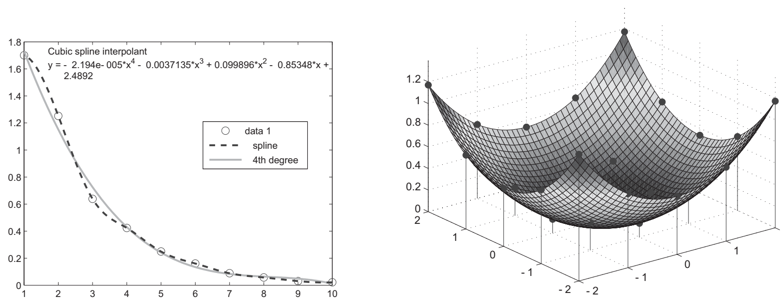


Figure 19.19 a) and b)

The Function `interp2` offers appropriate methods to interpolate the data given over a two dimensional rectangular grid (see 19.7.2.1, p. 998). To interpolate irregularly distributed data is served by calling `griddata`.

■ The command sequence

>> `[X, Y] = meshgrid(-2 : 1 : 2);    F = 4 - sqrt(16 - X.^2 - Y.^2);`

>> `[Xe, Ye] = meshgrid(-2 : 0.1 : 2);    S = interp2(X, Y, F, Xe, Ye, 'spline');`

>> `surf(Xe, Ye, S)`

>> `hold on;    stem3(X, Y, F, 'fill')`

realizes the bivariable cubic spline interpolation of the function $f(x, y) = \sqrt{16 - x^2 - y^2}$ given on a grid. The interpolation spline is evaluated on a finer rectangular grid. **Fig.19.19b** represents the interpolation function where the data points are also shown.

5. Numerical Integration

Numerical integration is available in Matlab by the procedures `quad` and `quadl`. Both procedures are based on the recursive application of interpolation quadratures with adaptive step-size selection. `quad` is based on the Simpson formula, and in `quadl` the Lobatto formulas of higher order are applied (see 19.3.2, p. 964). In the case of sufficiently smooth integrand and higher accuracy requirements `quadl` works more effectively than `quad`.

■ As the first example the approximation of the definite integral $I = \int_0^1 \frac{\sin x}{x} dx$ (Integral sine see 8.2.5,**1.** p. 513) is considered.

>> `format long;    [I, fwerte] = quad(@(x)(sin(x)./x), 0, 1)`

```
Warning:  Divide by zero.    > In @(x)(sin(x)./x)   In quad at 62
I = 0.94608307007653   fwerte = 14
>>   format long;   [I,fwerte] = quadl(@(x)(sin(x)./x),0,1)
Warning:  Divide by zero.    > In @(x)(sin(x)./x)   In quadl at 64
I = 0.94608307036718   fwerte = 19
```

Both procedures obviously recognize the discontinuity of the integrand at the left endpoint of the interval, but the approximated value of the integral can be obtained without any difficulty. Based on the results of the same example in 19.3.4.2, p. 968 the number of function evaluations seems to be high but it is determined for the adaptive recursion.

```
>>   format long;   [I,fwerte] = quad(@(x)(sin(x)./x),0,1,1e-14)
I = 0.94608307036718   fwerte = 258
>>   format long;   [I,fwerte] = quadl(@(x)(sin(x)./x),0,1,1e-14)
I = 0.94608307036718   fwerte = 19
```

(The warning messages are not repeated here.) The determination of the accuracy of 10^{-14}, which is given as a further argument (the default is a 10^{-6} tolerance), makes the advantages of `quadl` obvious in this case.

■ The definite integral $I = \int_{-1000}^{1000} e^{-x^2}dx$ is to be determined.

```
>>   format long;   [I,fwerte] = quad(@(x)(exp(-x.^2)),-1000,1000,1e-10)
I = 1.77245385094233   fwerte = 585
>>   format long;   [I,fwerte] = quadl(@(x)(exp(-x.^2)),-1000,1000,1e-10)
I = 1.77245385090571   fwerte = 768
```

The flat shape of the integrand in a very wide part of the integration interval and the relatively steep peak at $x = 0$ make `quad` better in this case.

6. Numerical Solution of Differential Equations

Matlab offers several procedures to determine the numerical solutions of initial value problems of systems of first order ordinary differential equations. A standard procedure is `ode45`, in which Runge-Kutta methods of 4-th and 5-th order are applied with adaptive step-size selection (see 19.4.1.2, p. 969). To achieve higher accuracy the program `ode113` is more effective with implemented linear multi-step methods of predictor-corrector type (see 19.4.1.4, p. 971). Besides there are procedures which are especially effective for stiff systems of differential equations (see 19.4.1.5,**4.**, p. 973).

■ To solve the problem $y' = \frac{1}{4}(x^2 + y^2)$, $y(0) = 0$, by the Runge-Kutta method (see 19.4.1.2, p. 970) in the interval $[0, 1]$ the input is

```
>>   [x,y] = ode45(@(x,y)((x.^2 + y.^2)./4),[0 1],0);   plot(x,y)
```

The shape of the resulted solution is represented in **Fig.19.20a**.

■ To solve the special Lorenz-system (see ■ in 17.2.4.3, p. 887)

$$x_1' = 10(x_2 - x_1), \quad x_2' = 28x_1 - x_2 - x_1x_3, \quad x_3' = x_1x_2 - \frac{8}{3}x_3$$

in the interval $0 \le t \le 50$ with initial conditions $x(0) = (0, 1, 0)^T$ the following commands are used:

```
>>   [t,x] = ode45(@(t,x)([10*(x(2)-x(1));
          28*x(1)-x(2)-x(1)*x(3);x(1)*x(2)-8*x(3)/3]),[0 50],[0;1;0]);
>>   plot(x(:,1),x(:,3))
```

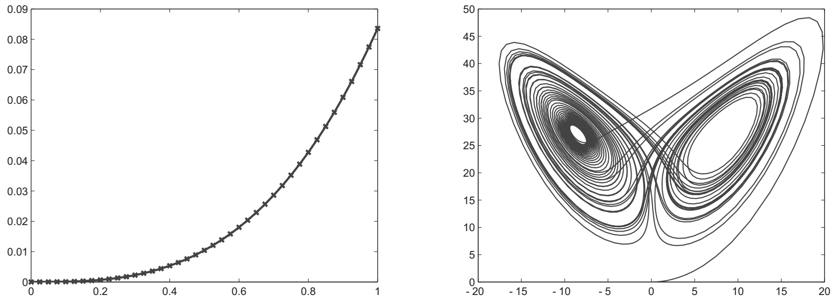


Figure 19.20 a) and b)

The last command creates a phase-diagram in the x_1, x_3 plane (see **Fig.19.20b**).

19.8.4.2 Mathematica

1. Tools for the Solution of Numerical Problems

The computer algebra system Mathematica offers a very effective tool that can be used to solve a large variety of numerical mathematical problems. The numerical procedures of Mathematica are totally different from symbolic calculations. Mathematica determines a table of values of the considered function according to certain previously given principles, similarly to the case of graphical representations, and it determines the solution using these values. Since the number of points must be finite, this could be a problem with " badly " behaving functions. Although Mathematica tries to choose more nodes in problematic regions, we have to suppose a certain continuity on the considered domain. This can be the cause of errors in the final result. It is advised to use as much information as possible about the problem under consideration, and if it is possible, then to perform calculations in symbolic form, even if this is possible only for subproblems.
In **Table 19.5**, we represent the operations for numerical computations:

Table 19.5 Numerical operations

`NIntegrate`	calculates definite integrals
`NSum`	calculates sums $\sum\limits_{i=1}^{n} f(i)$
`NProduct`	calculates products
`NSolve`	numerically solves algebraic equations
`NDSolve`	numerically solves differential equations

After starting Mathematica the " Prompt " *In[1]* := is shown; it indicates that the system is ready to except an input. Mathematica denotes the output of the corresponding result by *Out[1]*. In general: The text in the rows denoted by *In[n]* := is the input. The rows with the sign *Out[n]* are given back by Mathematica as answers. The arrow –> in the expressions means, e.g., replace x by the value a.

2. Curve Fitting and Interpolation

1. Curve Fitting Mathematica can perform the fitting of chosen functions to a set of data using the least squares method (see 6.2.5, p. 454ff.) and the approximation in mean to discrete problems (see 19.6.2.2, p. 986). The general instruction is:

$$\texttt{Fit}[\{y_1, y_2, \ldots\}, funkt, x]. \qquad (19.291)$$

Here the values y_i form the list of data, $funkt$ is the list of the chosen functions, by which the fitting should be performed, and x denotes the corresponding domain of the independent variables. If $funkt$ is chosen, e.g., as $\texttt{Table}[x\hat{}i, \{i, 0, n\}]$, then the fitting will be made by a polynomial of degree n.

■ Let the following list of data be given:

In[1] $:= l = \{1.70045, 1.2523, 0.638803, 0.423479, 0.249091, 0.160321, 0.0883432, 0.0570776, 0.0302744, 0.0212794\}$

With the input

In[2] $:= f1 = \texttt{Fit}[l, \{1, x, x\hat{}2, x\hat{}3, x\hat{}4\}, x]$

it is supposed that the elements of l are assigned to the values $1, 2, \ldots, 10$ of x. The result is the following approximation polynomial of degree four:

Out[2] $= 2.48918 - 0.853487x + 0.0998996x^2 - 0.00371393x^3 - 0.0000219224x^4$

With the command

In[3] $:= \texttt{Plot}[\texttt{ListPlot}[l, \{x, 10\}], f1, \{x, 1, 10\}, \texttt{AxesOrigin}{-}{>}\ \{0, 0\}]$

a representation of the data and the approximation curve given in **Fig. 19.21a** can be obtained.

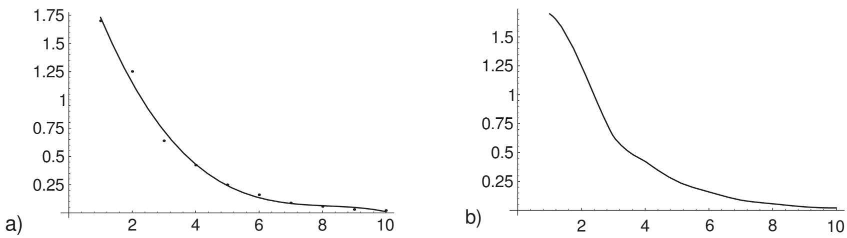


Figure 19.21

For the given data this is completely satisfactory. The terms are the first four terms of the series expansion of $e^{1-0.5x}$.

2. Interpolation Mathematica offers special algorithms for the determination of interpolation functions. They are represented as so-called `interpolating function` objects, which are formed similarly to pure functions. The directions for using them are in **Table 19.6**. Instead of the single function values y_i a list of function values and values of specified derivatives can be given at the given points.

■ With *In[3]* $:= \texttt{Plot}[\texttt{Interpolation}[l][x], \{x, 1, 10\}]$ one gets **Fig. 19.21b**. Obviously Mathematica gives a precise correspondence to the data list.

Table 19.6 Commands for interpolation

$\texttt{Interpolation}[\{y_1, y_2, \ldots\}]$	gives an approximation function with the values y_i for the values $x_i = i$ as integers
$\texttt{Interpolation}[\{\{x_1, y_1\}, \{x_2, y_2\}, \ldots\}]$	gives an approximation function for the point–sequence (x_i, y_i)

3. Numerical Solution of Polynomial Equations

As shown in 20.3.2.1, p. 1038 Mathematica can determine the roots of polynomials numerically. The command is $\texttt{NSolve}[p[x] == 0, x, n]$, where n prescribes the accuracy by which the calculations should be done. If n is omitted, then the calculations are made to machine accuracy. The complete solution is got, i.e., m roots, if the input polynomial is of degree m.

■ *In[1]* $:= \texttt{NSolve}[x\hat{}6 + 3x\hat{}2 - 5 == 0]$

Out[1] $= \{x{-}{>} -1.07432\}, \{x{-}{>} -0.867262 - 1.15292\texttt{I}\}, \{x{-}{>} -0.867262 + 1.15292\texttt{I}\},$

$\{x{-}{>}\ 0.867262 - 1.15292\texttt{I}\}, \{x{-}{>}\ 0.867262 + 1.15292\texttt{I}\}, \{x{-}{>}\ 1.07432\}\}.$

4. Numerical Integration

For numerical integration Mathematica offers the procedure `NIntegrate`. Differently from the symbolical method, here it works with a table of values of the integrand. Two improper integrals are considered (see 8.2.3, p. 506) as examples.

■ **A:** *In[1]* := `NIntegrate[Exp[`$-x$`^2], {`x`, -Infinity, Infinity}]` *Out[1]* = 1.77245.

■ **B:** *In[2]* := `NIntegrate[1/`x`^2, {`x`, -1, 1}]`

Power::infy: Infinite expression $\frac{1}{0}$ encountered.
NIntegrate::inum: Integrand ComplexInfinity is not numerical at$\{x\} = \{0\}$.

Mathematica recognizes the discontinuity of the integrand at $x = 0$ in example **B** and gives a warning. Mathematica applies a table of values with a higher number of nodes in the problematic domain, and it recognizes the pole. However, the answer can be still wrong.

Mathematica applies certain previously specified options for numerical integration, and in some special cases they are not sufficient. The minimal and the maximal number of recursion steps, by which Mathematica works in a problematic domain, can be determined with parameters `MinRecursion` and `MaxRecursion`. The default options are always 0 and 6. If these values are increased, then although Mathematica works slower, it gives a better result.

■ *In[3]* := `NIntegrate[Exp[`$-x$`^2], {`x`, -1000, 1000}]` Mathematica cannot find the peak at $x = 0$, since the integration domain is too large, and the answers is:

NIntegrate::ploss:
Numerical integration stopping due to loss of precision. Achieved neither the requested PrecisionGoal nor AccuracyGoal; suspect one of the following: highly oscillatory integrand or the true value of the integral is 0.
Out[3] $= 1.34946 \cdot 10^{-26}$

If the requirement is

In[4] := `NIntegrate[Exp[-x^2], {`x`, -1000, 1000}, MinRecursion-> 3, MaxRecursion-> 10]`,

then the result is

Out[4] = 1.77245

Similarly, a result closer to the actual value of the integral can be got with the command:

`NIntegrate[`$fun, \{x, x_a, x_1, x_2, \ldots, x_e\}$`]`. (19.292)

One can give the points of singularities x_i between the lower and upper limit of the integral to force Mathematica to evaluate more accurately.

5. Numerical Solution of Differential Equations

In the numerical solution of ordinary differential equations and also in the solution of systems of differential equations Mathematica represents the result by an `InterpolatingFunction`. It allows us to get the numerical values of the solution at any point of the given interval and also to sketch the graphical representation of the solution function. The most often used commands are represented in **Table 19.7**.

Table 19.7 Commands for numerical solution of differential equations

`NDSolve[`$dgl, y, \{x, x_a, x_e\}$`]`	computes the numerical solution of the differential equation in the domain between x_a and x_e
`InterpolatingFunction[`$liste$`][`x`]`	gives the solution at the point x
`Plot[Evaluate[`$y[x]$`/. `$lös$`]], {`x, x_a, x_e`}]`	scetches the graphical representation

■ Solution of a differential equation describing the motion of a heavy object in a medium with friction. The equations of motion in two dimensions are

$$\ddot{x} = -\gamma\sqrt{\dot{x}^2 + \dot{y}^2} \cdot \dot{x}, \quad \ddot{y} = -g - \gamma\sqrt{\dot{x}^2 + \dot{y}^2} \cdot \dot{y}.$$

The friction is supposed to be proportional to the velocity. If $g = 10, \gamma = 0.1$ are substituted, then the following command can be given to solve the equations of motion with initial values $x(0) = y(0) = 0$ and $\dot{x}(0) = 100, \dot{y}(0) = 200$:

In[1] := dg = `NDSolve`$[\{x''[t] == -0.1$`Sqrt`$[x'[t]$^$2 + y'[t]$^$2]\ x'[t], y''[t] == -10$

-0.1`Sqrt`$[x'[t]$^$2 + y'[t]$^$2]\ y'[t], x[0] == y[0] == 0, x'[0] == 100, y'[0] == 200\},$

$\{x, y\}, \{t, 15\}]$

Mathematica gives the answer by the interpolating function:

Out[1] = $\{\{x$–> `InterpolatingFunction`$[\{0., 15.\}, $<>$],$

y–> `InterpolatingFunction`$[\{0., 15.\}, $<>$]\}\}$

The solution

In[2] := `ParametricPlot`$[\{x[t], y[t]\}/.dg, \{t, 0, 2\},$ `PlotRange`–> `All`$]$

is represented as a parametric curve (**Fig. 19.22a**).

`NDSolve` accepts several options which affect the accuracy of the result.
The accuracy of the calculations can be given by the command `AccuracyGoal`. The command `PrecisionGoal` works similarly. During calculations, Mathematica works according to the so-called `WorkingPre cision`, which should be increased by five units in calculations requiring higher accuracy.

The numbers of steps by which Mathematica works in the considered domain is prescribed as 500. In general, Mathematica increases the number of nodes in the neighborhood of the problematic domain. In the neighborhood of singularities it can exhaust the step limit. In such cases, it is possible to increase the number of steps by `MaxSteps`. It is also possible the prescribe `Infinity` for `MaxSteps`.

■ The equations for the Foucault pendulum are:

$$\ddot{x}(t) + \omega^2 x(t) = 2\Omega\dot{y}(t), \quad \ddot{y}(t) + \omega^2 y(t) = -2\Omega\dot{x}(t).$$

With $\omega = 1,\ \Omega = 0.025$ and the initial conditions $x(0) = 0,\ y(0) = 10,\ \dot{x}(0) = \dot{y}(0) = 0$ the solution is:

In[3] := $dg3$ = `NDSolve`$[\{x''[t] == -x[t] + 0.05y'[t], y''[t] == -y[t] - 0.05x'[t],$

$x[0] == 0, y[0] == 10, x'[0] == y'[0] == 0\}, \{x, y\}, \{t, 0, 40\}]$

Out[3] = $\{\{x$–> `InterpolatingFunction`$[\{0., 40.\}, $<>$],$

y–> `InterpolatingFunction`$[\{0., 40.\}, $<>$]\}\}$

With

In[4] := `ParametricPlot`$[\{x[t], y[t]\}/.dg3, \{t, 0, 40\},$ `AspectRatio`–> $1]$

one gets **Fig. 19.22b**.

19.8.4.3 Maple

The computer algebra system Maple can solve several problems of numerical mathematics with the use of built-in approximation methods. The number of nodes, which is required by the calculations, can be determined by specifying the value of the global variable `Digits` as an arbitrary n. But it should be kept in mind that selecting a higher n than the prescribed value results in a lower speed of calculation.

1. Numerical Calculation of Expressions and Functions

After starting Maple, the symbol “ Prompt ” > is shown, which denotes the readyness for input. Connected in- and outputs are often represented in one row, separated by the *arrow operator* $\longrightarrow$.

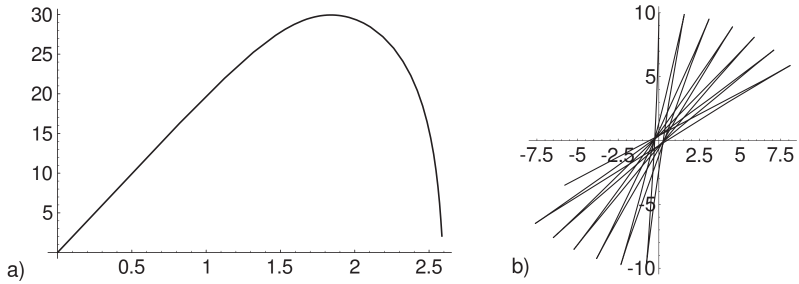


Figure 19.22

1. Operator `evalf` Numerical values of expressions containing built-in and user-defined functions which can be evaluated as a real number, can be calculated with the command

$$\texttt{evalf}(expr, n). \tag{19.293}$$

expr is the expression whose value should be determined; the argument n is optional, it is for evaluation to n digits accuracy. Default accuracy is set by the global variable `Digits`.

■ Prepare a table of values of the function $y = f(x) = \sqrt{x} + \ln x$.
First, the function is defined by the arrow operator:

$$> f := z \text{ -> } \texttt{sqrt}(z) + \ln(z); \longrightarrow f := z \text{ -> } \sqrt{x} + \ln x.$$

Then the required values of the function can be got with the command $\texttt{evalf}(f(x));$, where a numerical value should be substituted for x.

A table of values of the function with steps size 0.2 between 1 and 4 can be obtained by

$$> \texttt{for } x \texttt{ from } 1 \texttt{ by } 0.2 \texttt{ to } 4 \texttt{ do print}(f[x] = \texttt{evalf}(f(x), 12)) \texttt{ od};$$

Here, it is required to work with twelve digits.
Maple gives the result in the form of a one-column table with elements in the form $f_{[3.2]} = 2.95200519181$.

2. Operator `evalhf`*(expr)*: Beside `evalf` there is the operator `evalhf`. It can be used in a similar way to `evalf`. Its argument is also an expression which has a real value. It evaluates the symbolic expression numerically, using the hardware floating-point double-precision calculations available on the computer. A Maple floating-point value is returned. Using `evalhf` speeds up your calculations in most cases, but you lose the definiable accuracy of using `evalf` and `Digits` together. For instance in the problem in 19.8.2, p. 1004, it can produce a considerable error.

2. Numerical Solution of Equations

By using Maple equations or systems of equations can be solved numerically in many cases.
The command to do this is `fsolve`. It has the syntax

$$\texttt{fsolve}(eqn, var, option). \tag{19.294}$$

This command determines real solutions. If *eqn* is in polynomial form, the result is all the real roots. If *eqn* is not in polynomial form, it is likely that `fsolve` will return only one solution. The available options are given in **Table 19.8**.

Table 19.8 Options for the command `fsolve`

`complex`	determines a complex root (or all roots of a polynomial)
`maxsols` $= n$	determines at least the n roots (only for polynomial equations)
`fulldigits`	ensures that `fsolve` does not lower the number of digits used during computations
`intervall`	looks for roots in the given interval

■ **A:** Determination of all solutions of a polynomial equation $x^6 + 3x^2 - 5 = 0$. With

$> eq := x\hat{}6 + 3 * x\hat{}2 - 5 = 0:$

the result is

$>$ `fsolve`$(eq, x); \longrightarrow -1.074323739,\ 1.074323739$

Maple determined only the two real roots. With the option `complex`, also the complex roots are obtained:

$>$ `fsolve`$(eq, x,$ `complex`$);$

$-1.074323739, -0.8672620244 - 1.152922012\mathtt{I}, -0.8672620244 + 1.152922012\mathtt{I},$
$0.8672620244 - 1.152922012\mathtt{I}, 0.8672620244 + 1.152922012\mathtt{I}, 1.074323739$

■ **B:** Determination of both solutions of the transcendental equation $e^{-x^3} - 4x^2 = 0$.
After defining the equation

$> eq :=$ `exp`$(-x\hat{}3) - 4 * x\hat{}2 = 0$

the result is

$>$ `fsolve`$(eq, x); \longrightarrow 0.4740623572$

as the positive solution. With

$>$ `fsolve`$(eq, x, x = -2..0); \longrightarrow -0.5412548544$

Maple also determines the second (negative) root.

3. Numerical Integration

The determination of definite integrals is often possible only numerically. This is the case when the integrand is too complicated, or if the primitive function cannot be expressed by elementary functions. The command to determine a definite integral in Maple is `evalf`:

`evalf`$($`int`$(f(x), x = a..b), n).$ (19.295)

Maple calculates the integral by using an approximation formula.

■ Calculation of the definite integral $\int_{-2}^{2} \mathrm{e}^{-x^3}\,dx$. Since the primitive function is not known, for the integral command the following answer is got

$>$ `int`$($`exp`$(-x\hat{}3), x = -2..2); \longrightarrow \int_{-2}^{2} e^{-x^3}\,dx.$

If

$>$ `evalf`$($`int`$($`exp`$(-x\hat{}3), x = -2..2), 15);$

is typed,then the answer is 277.745841695583.

Maple used the built-in approximation method for numerical integration with 15 digits.

In certain cases this method fails, especially if the integration interval is too large. Then, another approximation procedure can be tried with the call to a library

```
readlib(`evalf/int`):
```

which applies an adaptive Newton method.

■ The input

$>$ `evalf`$($`int`$($`exp`$(-x\hat{}2), x = -1000..1000));$

results in an error message. With

```
> readlib(`evalf/int`):
> `evalf/int`(exp(-x^2), x = -1000..1000, 10, _NCrule);
  1.772453851
```

the correct result is obtained. The third argument specifies the accuracy and the last one specifies the internal notation of the approximation method.

4. Numerical Solution of Differential Equations

Ordinary differential equations can be solved with the Maple operation `dsolve`. However, in most cases it is not possible to determine the solution in closed form. In these cases, it can be solved numerically, where the corresponding initial conditions have to be given.
In order to do this, the command `dsolve` is used in the form

$$\texttt{dsolve}(deqn, var, \texttt{numeric}) \tag{19.296}$$

with the option `numeric` as a third argument. Here, the argument *deqn* contains the actual differential equation and the initial conditions. The result of this operation is a procedure, and if it is denoted, e.g., by f, for using the command $f(t)$, the value of the solution function at the value t of the independent variable is returned.
Maple applies the Runge-Kutta method to get this result (see 19.4.1.2, p. 969). The default accuracy for the relative and for the absolute error is $10^{-\text{Digits}+3}$. The user can modify these default error tolerances with the global symbols `_RELERR` and `_ABSERR`. If there are some problems during calculations, then Maple gives different error messages.

■ At solving the problem given in the Runge-Kutta methods in 19.4.1.2, p. 970, Maple gives:

$> r := \texttt{dsolve}(\{\texttt{diff}(y(x), x) = (1/4) * (x\hat{}2 + y(x)\hat{}2), y(0) = 0\}, y(x), \texttt{numeric});$

$r := \texttt{proc}$ \``dsolve/numeric/result2`\` $(x, 1592392, [1])$ `end`

With

$> \texttt{r}(0.5); \longrightarrow \{x(.5) = 0.5000000000, y(x)(.5) = 0.01041860472\}$

we can determine the value of the solution, e.g., at $x = 0.5$.

20 Computer Algebra Systems- Example Mathematica

20.1 Introduction

20.1.1 Brief Characterization of Computer Algebra Systems

20.1.1.1 General Purpose of Computer Algebra Systems

The development of computers has made possible the introduction of computer algebra systems for "doing mathematics". They are software systems able to perform mathematical operations formally. These systems, such as Macsyma, Reduce, Derive, Maple, Mathematica, Matlab, Sage, can also be used on relatively small computers (PC), and with their help, we can transform complicated expressions, calculate derivatives and integrals, solve systems of equations, represent functions of one and of several variables variables graphically, etc. They can *manipulate* mathematical expressions, i.e., they can transform and simplify mathematical expressions according to mathematical rules if this is possible in closed form. They also provide a wide range of numerical solutions to required accuracy, and they can represent functional dependence between data sets graphically.

Most computer algebra systems can import and export data. Besides a basic offer of definitions and procedures which are activated at every start of the system, most systems provide a large variety of libraries and program packages from special fields of mathematics, which can be loaded and activated on request (see [20.15],[20.16]). Computer algebra systems allow users to build up their own packages [20.11]–[20.14].

However, the possibilities of computer algebra systems should not be overestimated. They spare us the trouble of boring, time-demanding, and mechanical computations and transformations, but they do not save us from thinking.
For frequent errors see 19.8.2, p. 1004.

20.1.1.2 Restriction to Mathematica

The systems are under perpetual developing. Therefore, every concrete representation reflects only a momentary state. Here we introduce the basic idea and applications of these systems for the most important fields of mathematics. This introduction will help for the first steps in working with computer algebra systems. In particular, we discuss Mathematica compatible until Version 10. This system seems to be very popular among users, and the basic structures of the other systems are similar.

In this book, we do not discuss how computer algebra systems are installed on computers. It is assumed that the computer algebra system has already been started by a command, and it is ready to communicate by command lines or in a Windows-like graphical environment.

The input and output is always represented for Mathematica (see 19.8.4.2, **1.**, p. 1016) in rows which are distinguished from other text parts, e.g., in the form

In[1] := `Solve`$[3\ x - 5 == 0, x]$ (20.1)

System specific symbols (commands, type notation, etc.) will be represented in typewriter style.

In order to save space, we often write the input and the output in the same row in this book, and we separate them by the symbol $\longrightarrow$.

20.1.1.3 Two Introducing Examples of Basic Application Fields

1. Manipulation of Formulas

Formula manipulation means here the transformation of mathematical expressions in the widest sense, e.g., simplification or transformation into a useful form, representation of the solution of equations or systems of equations by algebraic expressions, differentiation of functions or determination of indefinite integrals, solution of differential equations, formation of infinite series, etc.

I.N. Bronshtein et al., *Handbook of Mathematics*,
DOI 10.1007/978-3-662-46221-8_20

■ Solution of the following quadratic equation:

$$x^2 + ax + b = 0 \quad \text{with} \quad a, b \in \mathbb{R}. \tag{20.2a}$$

In Mathematica, one types:

$$\texttt{Solve}[x^2 + a\ x + b == 0, x]\,. \tag{20.2b}$$

After pressing the corresponding input key/keys (ENTER or SHIFT+RETURN, depending on the operation system), Mathematica replaces this row by

$$\textit{In[1]} := \texttt{Solve}[x^2 + a\ x + b == 0, x] \tag{20.2c}$$

and starts the evaluation process. In a moment, the answer appears in a new row

$$\textit{Out[1]} = \{\{x \text{->} \frac{1}{2}\left(-a - \sqrt{a^2 - 4b}\right)\}, \{x \text{->} \frac{1}{2}\left(-a + \sqrt{a^2 - 4b}\right)\}\}. \tag{20.2d}$$

Mathematica has solved the equation and both solutions are represented in the form of a *list* consisting of two sublists.

2. Numerical Calculations

Computer algebra systems provide many procedures to handle numerical problems of mathematics. These are solutions of algebraic equations, linear systems of equations, the solutions of transcendental equations, calculation of definite integrals, numerical solutions of differential equations, interpolation problems, etc.

■ Problem: Solution of the equation

$$x^6 - 2x^5 - 30x^4 + 36x^3 + 190x^2 - 36x - 150 = 0. \tag{20.3a}$$

Although this equation of degree six cannot be solved in closed form, it has six real roots, which are to be determined numerically.

In Mathematica the input is:

$$\textit{In[1]} := \texttt{NSolve}[x^6 - 2x^5 - 30x^4 + 36x^3 + 190x^2 - 36x - 150 == 0, x] \tag{20.3b}$$

It results in the answer:

$$\textit{Out[1]} = \{\{x \text{->} -4.42228\}, \{x \text{->} -2.14285\}, \{x \text{->} -0.937347\}, \{x \text{->} 0.972291\},$$
$$\{x \text{->} 3.35802\}, \{x \text{->} 5.17217\}\} \tag{20.3c}$$

This is a list of the six solutions with a certain accuracy which will be discussed later.

20.2 Important Structure Elements of Mathematica

Mathematica is a computer algebra system, developed by Wolfram Research Inc. A detailed description of Mathematica can be found in [20.11]–[20.16]. For the current version 10 see the Virtual Book in the online Help.

20.2.1 Basic Structure Elements of Mathematica

In Mathematica the basic structure elements are called *expressions*. Their syntax is (it is emphasized again, that the current objects are given by their corresponding symbol, by their names):

$$\texttt{obj}_0[\texttt{obj}_1, \texttt{obj}_2, \ldots, \texttt{obj}_n] \tag{20.4}$$

$\texttt{obj}_0$ is called the head of the expression; the number 0 is assigned to it. The parts $\texttt{obj}_i$ $(i = 1, \ldots, n)$ are the *elements* or *arguments* of the expression, and one can refer to them by their numbers $1, \ldots, n$.
In many cases the head of the expression is an operator or a function, the elements are the operands or variables on which the head acts.
Also the head, as an element of an expression, can be an expression, too. Square brackets are reserved in Mathematica for the representation of an expression, and they can be applied only in this relation.

■ The term $x\hat{}2+2*x+1$, which can also be entered in this infix form (and also in the nicer, preferred form x^2+2x+1) in Mathematica, has the complete form (`FullForm`)

$$\texttt{Plus}[1, \texttt{Times}[2, x], \texttt{Power}[x, 2]]$$

which is also an expression. `Plus`, `Power` and `Times` denote the corresponding arithmetical operations.

The example shows that all simple mathematical operators exist in prefix form in the internal representation, and the term notation is only a facility in Mathematica.
Parts of expressions can be separated. This can be done with `Part`[*expr*, *i*], where i is the number of the corresponding element. In particular, $i = 0$ gives back the head of the expression.

■ If entering in Mathematica

In[1] $:= x^2 + 2x + 1,$

then after the SHIFT and ENTER keys together are pressed, Mathematica answers

Out[1] $= 1 + 2x + x^2$

Mathematica analyzed the input and returned it in mathematical standard form. If the input had been terminated by a semicolon, then the output would have been suppressed.
If entering

In[2] := `FullForm`[%]

then the answer is

Out[2] = `Plus`[1, `Times`[2, x], `Power`[x, 2]]

The sign % in the square brackets tells Mathematica that the argument of this input is the last output. From this expression it is possible to get, e.g., the third element

In[3] := `Part`[%, 3] as *Out[3]* $= \texttt{x}^2$

which is again an expression in this case.

Symbols in Mathematica are the notation of the basic objects; they can be any sequence of letters and numbers but they must not begin with a number. The special sign $ is also allowed. Upper-case and lower-case letters are distinguished. Reserved symbols begin either with a capital letter, or with the sign $, and in compound words also the second word begins with a capital letter, if it has a separate meaning. Users are advised to create their own symbols starting with lower-case letters.

20.2.2 Types of Numbers in Mathematica

20.2.2.1 Basic Types of Numbers

Mathematica knows four types of numbers represented in **Table 20.1**.

Table 20.1 Types of numbers in Mathematica

Type of number	Head	Characteristic	Input
Integers	`Integer`	exact integer, arbitrarily long	*nnnnn*
Rational numbers	`Rational`	fraction of coprimes in form `Integer/Integer`	*pppp/qqqq*
Real numbers	`Real`	floating-point number, arbitrary given precision	*nnnn.mmmm*
Complex numbers	`Complex`	complex number in the form *number+number* ∗`I`	

Real numbers, i.e., floating-point numbers, can be arbitrarily long. If an integer *nnn* is written in the form *nnn.*, then Mathematica considers it as a floating-point number, that is, of type `Real`.
The type of a number x can be determined with the command `Head`[x]. Hence, *In[1]* := `Head`[51] results in *Out[1]* = `Integer`, while *In[2]* := `Head`[51.] *Out[2]* = `Real`. The real and imaginary components of a complex number can belong to any type of numbers. A number such as $5.731 + 0\,\texttt{I}$ is considered as a `Real` type by Mathematica, while $5.731 + 0.$ I is of type `Complex`, since 0. is considered as a floating-point approximation of 0.

There are some further operations, which give information about numbers. So,

$$In[3] := \texttt{NumberQ}[51] \text{ results in } Out[3] = \texttt{True}, \quad (20.5a)$$

Otherwise, if x is manifestly not a number, as e.g. $x = \pi$, then the output is *Out[3]* =`False`. However, `NumericQ[`π`]` gives `True`. Here, `True` and `False` are the symbols for Boolean constants.
`IntegerQ`$[x]$ tests if x is an integer, or not, so

$$In[4] := \texttt{IntegerQ}[2.] \longrightarrow Out[4] = \texttt{False} \quad (20.5b)$$

Similar tests can be performed for numbers with heads `EvenQ`, `OddQ` and `PrimeQ`. Their meanings are obvious. So, one gets

$$In[5] := \texttt{PrimeQ}[1075643] \longrightarrow Out[5] = \texttt{True} \quad (20.5c)$$

while

$$In[6] := \texttt{PrimeQ}[1075641] \longrightarrow Out[6] = \texttt{False} \quad (20.5d)$$

These last tests belong to a group of test operators, called predicates or criteria, which all end by `Q` and always answer `True` or `False` in the sense of a logical test (including a type check).

20.2.2.2 Special Numbers

In Mathematica, there are some special numbers which are often needed, and they can be called with arbitrary accuracy. They include π with the symbol `Pi`, e with the symbol `E`, $\frac{\pi}{180^\circ}$ as the transformation factor from degree measure into radian measure with the constant `Degree`, `Infinity` as the symbol for ∞ and the imaginary unit `I`.

20.2.2.3 Representation and Conversion of Numbers

Numbers can be represented in different forms which can be converted into each other. So, every real number x can be represented by a floating-point number `N`$[x,\ n]$ with an n-digit precision.

$$IN[1] := \texttt{N}[E,\ 20] \quad \text{yields} \longrightarrow Out[1] = 2.7182818284590452354 \quad (20.6a)$$

With `Rationalize`$[x,\ dx]$, the number x with an accuracy dx can be converted into a rational number, i.e., into the fraction of two integers.

$$In[2] := \texttt{Rationalize}[\%,\ 10^{-}5] \longrightarrow Out[2] = \frac{1071}{394} \quad (20.6b)$$

With 0 accuracy, Mathematica gives the possible best approximation of the number x by a rational number.
Numbers of different number systems can be converted into each other. With `BaseForm`$[x,\ b]$, the number x given in the decimal system is converted into the corresponding number in the number system with base $b \le 36$. If $b > 10$, then the consecutive letters of the alphabet $a, b, c, \ldots$ are used for the further digits having a meaning greater than ten.

■ **A**: $In[1] := \texttt{BaseForm}[255,\ 16] \longrightarrow Out[1] = ff_1$ (20.7a)

$In[2] := \texttt{BaseForm}[\texttt{N}[E,\ 10],\ 8] \longrightarrow Out[2] = 2.5576052131_8$ (20.7b)

The reversed transformation can be performed by $b^{\wedge\wedge}mmmm$.

■ **B**: $In[1] := 8^{\wedge\wedge}735 \longrightarrow Out[1] = 477$ (20.7c)

Numbers can be represented with arbitrary precision (the default here is the hardware precision), and for large numbers so-called scientific form is used, i.e., the form $n.mmmm10^{\wedge} \pm qq$.

20.2.3 Important Operators

Several basic operators can be written in infix form (as in the classical form in mathematics) $<\ symb_1\ op\ symb_2\ >$. However, in every case, the complete form of this simplified notation is the expression $op[symb_1,\ symb_2]$. The most often occurring operators and their complete form are collected in **Table 20.2**. Most symbols in **Table 20.2** are obvious. For multiplication in the form $a\,b$, the space between the factors is very important.

Table 20.2 Important Operators in Mathematica

$a + b$	Plus[a, b]	$u == v$	Equal[u, v]
$a\,b$ or $a * b$	Times[a, b]	$w\,!= v$	Unequal[w, v]
a^b ore a^b	Power[a, b]	$r > t$	Greater[r, t]
a/b	Times[a, Power[b, -1]]	$r \geq t$	GreaterEqual[r, t]
u-> v	Rule[u, v]	$s < t$	Less[s, t]
$r = s$	Set[r, s]	$s \leq t$	LessEqual[s, t]

The expressions with the heads Rule and Set will be explained. Set assigns the value of the expression s on the right-hand side, e.g., a number, to the expression r on the left-hand side, e.g., a variable. From here on, r is represented by this value until this assignment is changed. The change can be done either by a new assignment or by $x = .$ or by Clear[x], i.e., by releasing every assignment so far. The construction Rule should be considered as a transformation rule. It occurs together with /. which is the substitution operator.

Replace[t, u -> v] or t/. u-> v means that every occurance u in the expression t will be replaced by the expression v.

■ *In[1]* := $x + y^2$ /. y-> $a + b$ $\longrightarrow$ *Out[1]* = $(a + b)^2 + x$

It is typical in the case of both operators that the right-hand side is evaluated immediately after the assignment or transformation rule. So, the left-hand side will be replaced by this evaluated right-hand side at every later call.

Here, two further operators have to be mentioned with delayed evaluation.

The FullForm of $u := v$ is SetDelayed[u, v] and (20.8a)

the FullForm of $u : -> v$ is RuleDelayed[u, v] (20.8b)

The assignment or the transformation rule are also valid here until it is changed. Although the left-hand side is always replaced by the right side, the right-hand side is evaluated for the first time only at the moment when the left one is called.

The expression $u == v$ or Equal[u, v] returns True if u and v are identical. Equal is used, e.g., in manipulation of equations.

20.2.4 Lists

20.2.4.1 Notions

Lists are important tools in Mathematica for the manipulation of whole groups of quantities, which are important in higher-dimensional algebra and analysis.

A *list* is a collection of several objects into a new object. In the list, each object is distinguished only by its place in the list. The construction of a list is made either by the command

List[$a1$, $a2$, $a3$, …] or by {$a1$, $a2$, $a3$, …} (20.9)

if the elements can be simply enumerated. To explain the work with lists, a particular list is used, denoted by $l1$:

In[1] := $l1$ = List[$a1$, $a2$, $a3$, $a4$, $a5$, $a6$] $\longrightarrow$ *Out[1]* = {$a1$, $a2$, $a3$, $a4$, $a5$, $a6$} (20.10)

Mathematica applies a short form to the output of the list: It is enclosed in curly braces.

Table 20.3 represents commands which choose one or more elements from a list, and the output is a "sublist".

■ For the list $l1$ in (20.9) one gets, e.g.,

In[2] := First[$l1$] $\longrightarrow$ *Out[2]* = $a1$ *In[3]* := $l1$[[3]] $\longrightarrow$ *Out[3]* = $a3$

In[4] := $l1$[[{2, 4, 6}]] $\longrightarrow$ *Out[4]* = {$a2$, $a4$, $a6$}

In[5] := Take[$l1$, 2] $\longrightarrow$ *Out[5]* = {$a1$, $a2$}

Table 20.3 Commands for the choice of list elements

`First`[l], `Last`[l]	selects the first/last element
`Most`[l], `Rest`[l]	selects the elements except the last/first one
`Part`[l, n] or l[[n]]	selects the n-th element
`Part`[l, {$n1$, $n2$, ...}]	gives a list of the elements with the given numbers
l[[{$n1$, $n2$, ...}]]	equivalent to the previous operation
`Take`[l, m]	gives the list of the first m elements of l
`Take`[l, {m, n}]	gives the list of the elements from m through n
`Drop`[l, n]	gives the list without the first n elements
`Drop`[l, {m, n}]	gives the list without the elements from m through n

20.2.4.2 Nested Lists

The elements of lists can also be lists, so nested lists can be obtained. If entering, e.g., for the elements of the previous list $l1$ (20.10)

In[1] $:= a1 = \{b11, b12, b13, b14, b15\}$

In[2] $:= a2 = \{b21, b22, b23, b24, b25\}$

In[3] $:= a3 = \{b31, b32, b33, b34, b35\}$

and analogously for $a4$, $a5$ and $a6$, then because of (20.10) a nested list (an array) is obtained which is not shown here explicitly. One can refer to the j-th element of the i-th sublist with the command `Part`[l, i, j]. The expression l[[i, j]] has the same result. In the above example (p. 1027), e.g.,

In[4] $:= l1[[3, 4]]$ yields *Out[4]* $= b34$

Furthermore, `Part`[l, {$i1$, $i2$ ...}, {$j1$, $j2$ ...}] or l[[{$i1$, $i2$, ...}, {$j1$, $j2$, ...}]] results in a list consisting of the elements numbered with $j1, j2 \ldots$ from the lists numbered with $i1, i2, \ldots$.

■ For the above example in 20.2.4.1, p. 1027

In[1] $:= l1[[\{3, 5\}, \{2, 3, 4\}]] \longrightarrow$ *Out[1]* $= \{\{b32, b33, b34\}, \{b52, b53, b54\}\}$

The idea of nesting lists is obvious from these examples. It is easy to create lists of three or higher dimensions, and it is easy to refer to the corresponding elements.

20.2.4.3 Operations with Lists

Mathematica provides several further operations by which lists can be monitored, enlarged or shortened (**Table 20.4**).

Table 20.4 Operations with lists

`Position`[l, a]	gives a list of the positions where a occurs in the list
`MemberQ`[l, a]	checks whether a is an element of the list
`Select`[l, *crit*]	picks out all elements of the list for which crit holds
`Cases`[l, *pattern*]	gives a list of elements which match the pattern
`FreeQ`[l, a]	checks if a does not occur in the list
`Prepend`[l, a]	changes the list by adding a to the front
`Append`[l, a]	changes the list by appending a to the end
`Insert`[l, a, i]	inserts a at position i in the list
`Delete`[l, {i, j, ...}]	delete the elements at positions $i, j, \ldots$ from the list
`ReplacePart`[l, a, i]	replace the element at position i by a

■ With `Delete`, the original list $l1$ (20.10) can be shortened by the term $a6$:

In[1] $:= l2 =$ `Delete`[$l1$, 6] $\longrightarrow$ *Out[1]* $= \{a1, a2, a3, a4, a5\}$,

where in the output the ai are shown by their values – they are lists themselves.

20.2.4.4 Tables

In Mathematica, several operations are available to create lists. One of them, which often occurs in working with mathematical functions, is the command `Table` shown in **Table 20.5**.

Table 20.5 Operation `Table`

`Table`[f, {$imax$}]	creates a list with $imax$ values of f: $f(1), f(2), \ldots, f(imax)$
`Table`[f, {i, $imin$, $imax$}]	creates a list with values of f from $imin$ to $imax$
`Table`[f, {i, $imin$, $imax$, di}]	the same as the last one, but by steps di

■ Table of binomial coefficients for $n = 7$:

In[1] := `Table`[`Binomial`[7, i], {i, 0, 7}]] $\longrightarrow$ *Out[1]* = {1, 7, 21, 35, 35, 21, 7, 1}

With `Table`, also higher-dimensional arrays can be created. With the expression

`Table`[f, {i, $i1$, $i2$}, {j, $j1$, $j2$}, ...]

a higher-dimensional, multiple nested table is obtained, i.e., entering

In[2] := `Table`[`Binomial`[i, j], {i, 1, 7}, {j, 0, i}]

the binomial coefficients are got up to degree 7:

Out[2] = {{1, 1}, {1, 2, 1}, {1, 3, 3, 1}, {1, 4, 6, 4, 1},

{1, 5, 10, 10, 5, 1}, {1, 6, 15, 20, 15, 6, 1}, {1, 7, 21, 35, 35, 21, 7, 1}}

The operation `Range` produces a list of consecutive numbers or equally spaced numbers:

`Range`[n] yields the list {1, 2, ..., n}

Similarly, `Range`[$n1$, $n2$] and `Range`[$n1$, $n2$, dn] produce lists of numbers (arithmetic sequences) from $n1$ to $n2$ with step-size 1 or dn respectively. The command `Array` uses functions (as opposed to function values used by `Table`) to create lists. `Array[Exp,5]` yields $\{e, e^2, e^3, e^4, e^5\}$.

20.2.5 Vectors and Matrices as Lists

20.2.5.1 Creating Appropriate Lists

Several special (list) commands are available for defining vectors and matrices. A one-dimensional list of the form

$$v = \{v1,\ v2,\ \ldots,\ vn\} \qquad (20.11)$$

can always be considered as a vector in n-dimensional space with components $v1, v2, \ldots, vn$. The special operation `Array`[v, n] produces the list (the vector) $\{v[1], v[2], \ldots, v[n]\}$. Symbolic vector operations can be performed with vectors defined in this way.

The two-dimensional lists $l1$ (see 20.2.4.2, p. 1028) and $l2$ (see 20.2.4.3, p. 1028) can be considered as matrices with rows i and columns j. In this case bij would be the element of the matrix in the i-th row and the j-th column. A rectangular matrix of type (6,5) is defined by $l1$, and a square matrix of type $(5, 5)$ by $l2$.

With the operation `Array`[b, {n, m}] a matrix of type (n, m) is generated, whose elements are denoted by $b[i,\ j]$. The rows are numbered by i, i changes from 1 to n; by j the columns are numbered from 1 to m. In this symbolic form $l1$ can be created as

$$l1 = \texttt{Array}[b,\ \{6,\ 5\}], \quad (20.12a) \quad \text{where } b[i,\ j] = bij \quad (i = 1, \ldots, 6;\ \ j = 1, \ldots, 5). \quad (20.12b)$$

Summarizing, lists can be created either by enumeration or by using the functions `Array, Range, Table`. Note that lists are different from sets in mathematics.

The operation `IdentityMatrix`[n] produces the n-dimensional unit matrix.

With the operation `DiagonalMatrix`[$list$] a diagonal matrix is produced with the elements of the $list$ in its main diagonal.

The operation `Dimension`[*list*] gives the size (number of rows, columns, ...) of a matrix, whose structure is given by a *list*. Finally, with the command `MatrixForm`[*list*], one gets a matrix-type representation of the *list*. A further possibility to define matrices is the following: Let $f(i, j)$ be a function of integers i and j. Then, the operation `Table`[$f[i\,j]$, $\{i,\ n\}$, $\{j,\ m\}$] defines a matrix of type (n, m), whose elements are the corresponding $f(i, j)$.

20.2.5.2 Operations with Matrices and Vectors

Mathematica allows formal manipulation of matrices and vectors. The operations given in **Table 20.6** can be applied.

Table 20.6 Operations with matrices

$c\,a$	matrix a is multiplied by the scalar c
$a\,.\,b$	the product of matrices a and b
`Det`[a]	the determinant of matrix a
`Inverse`[a]	the inverse of matrix a
`Transpose`[a]	the transpose of matrix a
`MatrixExp`[a]	the exponential function of matrix a
`MatrixPower`[a, n]	the n-th power of matrix a
`Eigenvalues`[a]	the eigenvalues of matrix a
`Eigenvectors`[a]	the eigenvectors of matrix a

■ **A** : *In[1]* := r = `Array`[a, {4, 4}] $\longrightarrow$ *Out[1]* = { { a[1, 1], a[1, 2], a[1, 3], a[1, 4] },
{ a[2, 1], a[2, 2], a[2, 3], a[2, 4] },
{ a[3, 1], a[3, 2], a[3, 3], a[3, 4] },
{ a[4, 1], a[4, 2], a[4, 3], a[4, 4] }}

In[2] := `Transpose`[r] $\longrightarrow$ *Out[2]* = { { a[1, 1], a[2, 1], a[3, 1], a[4, 1] },
{ a[1, 2], a[2, 2], a[3, 2], a[4, 2] },
{ a[1, 3], a[2, 3], a[3, 3], a[4, 3] },
{ a[1, 4], a[2, 4], a[3, 4], a[4, 4] }}

Here, the transpose matrix $r^{\top}$ of r is produced.

Let the general four-dimensional vector v be defined by

In[3] := v = `Array`[u, 4] $\longrightarrow$ *Out[3]* = {u[1], u[2], u[3], u[4]}

Now, the product of the matrix r and the vector v is again a vector (see Calculations with Matrices, 4.1.4, p. 272).

In[4] := $r.\,v$ $\longrightarrow$ *Out [4]* = { a[1, 1] u[1] + a[1, 2] u[2] + a[1, 3] u[3] + a[1, 4] u[4] ,
a[2, 1] u[1] + a[2, 2] u[2] + a[2, 3] u[3] + a[2, 4] u[4] ,
a[3, 1] u[1] + a[3, 2] u[2] + a[3, 3] u[3] + a[3, 4] u[4] ,
a[4, 1] u[1] + a[4, 2] u[2] + a[4, 3] u[3] + a[4, 4] u[4] }.

There is no difference between row and column vectors in Mathematica. In general, matrix multiplication is not commutative (see Calculations with Matrices 4.1.4, p. 272). The expression $r.\,v$ corresponds to the product in linear algebra when a matrix is multiplied by a column vector from the right, while $v.\,r$ means a multiplication by a row vector from the left.

■ **B:** In the section on Cramer's rule (4.5.2.3, p. 311) the linear system of equations $pt = b$ is solved with the matrix

In[1] := `MatrixForm`[p = {{2, 1, 3}, {1, −2, 1}, {3, 2, 2}}] $\longrightarrow$ *Out[1]* = $\begin{pmatrix} 2 & 1 & 3 \\ 1 & -2 & 1 \\ 3 & 2 & 2 \end{pmatrix}$

and vectors

In[2] := t = `Array`[x, 3] $\longrightarrow$ *Out[2]* = {x[1], x[2], x[3]}

In[3] := b = {9, −2, 7} $\longrightarrow$ *Out[3]* = {9, −2, 7}.

Since in this case Det[p]==13 $\neq 0$ holds, the system can be solved by $t = p^{-1}b$. This can be done by

In[4] := Inverse$[p].b$ with the output of the solution vector *Out[4]* = $\{-1, 2, 3\}$.

Note that a b calculates the componentwise product and Exp[a] gives the matrix containing the exp of the components of the matrix a.

20.2.6 Functions

20.2.6.1 Standard Functions

Mathematica knows several standard mathematical functions, which are listed in **Table 20.7**.

Table 20.7 A few standard functions

Exponential function	Exp[x]
Logarithmic functions	Log[x], Log[b,x]
Trigonometric functions	Sin[x], Cos[x], Tan[x], Cot[x], Sec[x], Csc[x]
Arc functions	ArcSin[x], ArcCos[x], ArcTan[x], ArcCot[x], ArcSec[x], ArcCsc[x]
Hyperbolic functions	Sinh[x], Cosh[x], Tanh[x], Coth[x], Sech[x], Csch[x]
Area functions	ArcSinh[x], ArcCosh[x], ArcTanh[x], ArcCoth[x], ArcSech[x], ArcCsch[x]

All these functions can be applied even with complex arguments.

In every case must be considered the single-valuedness of the functions. For real functions one branch of the function has to be chosen (if it is needed); for functions with complex arguments the principal value (see 14.5, p. 758) should be chosen.

20.2.6.2 Special Functions

Mathematica knows several special functions, which are not standard functions. **Table 20.8** lists some of these functions.

Table 20.8 Special functions

Bessel functions $J_n(z)$ and $Y_n(z)$	BesselJ[n,z], BesselY[n,z]
Modified Bessel functions $I_n(z)$ and $K_n(z)$	BesselI[n,z], BesselK[n,z]
Legendre polynomials $P_n(x)$	LegendreP[n,x]
Spherical harmonic $Y_l^m(\vartheta, \phi)$	SphericalHarmonicY[l, m, θ, ϕ]

Further functions can be loaded with the corresponding special packages of Mathematica

20.2.6.3 Pure Functions

Mathematica supports the use of so-called pure functions. A pure function is an anonymous function, an operation with no name assigned to it. They are denoted by Function[x, *body*]. The first argument specifies the formal parameters and the second one is the body of the function, i.e., *body* is an expression for the function of the variable x.

In[1] := Function[$x,\ x^3 + x^2$] $\longrightarrow$ *Out[1]* = Function[$x,\ x^3 + x^2$] (20.13)

and so

In[2] := Function[$x,\ x^3 + x^2$][c] gives *Out[2]* = $c^2 + c^3$. (20.14)

We can use a simplified version of this command. It has the form *body* &, where the variable is denoted by # . Instead of the previous two rows one can also write

In[3] := (#^{3} + #^{2}) & [c] *Out[3]* = $c^2 + c^3$. (20.15)

It is also possible to define pure functions of several variables:

Function[$\{x_1, x_2, \ldots\}$, *body*] or in short form *body* &, where the variables in *body* are denoted by the elements #1, #2, ... The sign & is very important for closing the expression, since it can be seen from this sign that the previous expression should be considered as a pure function. Let us remark that the

pure function #& is nothing else than the identity function: to any argument x it assigns x. Similarly, #1& corresponds to the projection onto the first coordinate axis.

20.2.7 Patterns

Mathematica allows users to define their own functions and to use them in calculations. With the command

$$\textit{In[1]} := \texttt{f}[x_] := \text{Polynomial}[x] \tag{20.16}$$

with Polynomial(x) as an arbitrary polynomial of variable x, a special function is defined by the user.

In the definition of the function f, there is no simple x, but $x_$ (pronounced x-blank) with a symbol _ for the blank. The symbol $x_$ means "something with the name x". From here on, every time when the expression f[*something*] occurs, Mathematica replaces it by its definition given above. This type of definition is called a *pattern*. The symbol *blank_* denotes the basic element of a pattern; $y_$ stands for y as a pattern. It is also possible to apply in the corresponding definition only a "_", that is $y\hat{}_$. This pattern stands for an arbitrary power of y with any exponent, thus, for an entire class of expressions with the same structure.

The essence of a pattern is that it defines a *structure*. When Mathematica checks an expression with respect to a pattern, it compares the structure of the elements of the expression to the elements of the pattern, Mathematica does not check mathematical equality! This is important in the following example: Let l be the list

$$\textit{In[2]} := l = \{1,\ y,\ y^a,\ y^{\sqrt{x}},\ \{\texttt{f}[y^{r/q}],\ 2^y\}\} \tag{20.17}$$

If one writes

$$\textit{In[3]} := l\ /.\ y\hat{}_ \;\text{->}\; yes \tag{20.18}$$

then Mathematica returns the list

$$\textit{Out[3]} = \{1,\ y,\ yes,\ yes,\ \{\texttt{f}[yes],\ 2^y\}\} \tag{20.19}$$

Mathematica checked the elements of the list with respect to its structural identity to its pattern $y\hat{}_$ and in every case when it determined coincidence it replaced the corresponding element by *yes*. The elements 1 and y were not replaced, since they have not the given structure, even though $y^0 = 1, y^1 = y$ holds.

Remark: Pattern comparison always happens in FullForm. If

$$\textit{In[4]} := b/y\ /.\ y\hat{}_ \;\text{->}\; yes \quad \text{is examined then} \quad \textit{Out[4]} = b\ yes \tag{20.20}$$

This is a consequence of the fact that FullForm of b/y is Times[b, Power[y, -1]], and for structure comparison the second argument of Times is identified as the structure of the pattern.

With the definition

$$\textit{In[5]} := \texttt{f}[x_] := x^3 \tag{20.21a}$$

Mathematica replaces, corresponding to the given pattern,

$$\textit{In[6]} = \texttt{f}[r] \quad \text{by} \quad \textit{Out[6]} = r^3 \quad \text{etc.} \tag{20.21b}$$

$$\textit{In[7]} := \texttt{f}[a] + \texttt{f}[x] \quad \text{yields} \quad \textit{Out[7]} = a^3 + x^3 \tag{20.21c}$$

If however,

$$\textit{In[8]} := \texttt{f}[x] := x^3, \quad \text{then for the same input} \quad \textit{In[9]} := \texttt{f}[a] + \texttt{f}[x] \tag{20.21d}$$

the output would be

$$\textit{Out[9]} = \texttt{f}[a] + x^3 \tag{20.21e}$$

In this case only the (fixed, single) input x corresponds to the definition.

20.2.8 Functional Operations

Functions operate on numbers and expressions. Mathematica can also perform operations with functions, since the names of functions are handled as expressions so they can be manipulated as expressions.

1. Inverse Function, Inverse Series The determination of the inverse function of a given function f can be made by the functional operation `InverseFunction` or `InverseSeries`.

■ **A:** *In[1]* := `InverseFunction`$[f]\,[x]$ $\longrightarrow$ *Out[1]* $= f^{-1}\,[x]$

■ **B:** *In[1]* := `InverseFunction[Exp]` $\longrightarrow$ *Out[1]* = `Log`

■ **C:** *In[1]* := `InverseSeries[Series`$[g[x], \{x, 0, 2\}]]$

$$\textit{Out[1]} = \frac{x - g[0]}{g'[0]} - \frac{g''[0](x - g[0])^2}{2g'[0]^3} + O[x - g[0]]^3$$

2. Differentiation Mathematica uses the possibility that the differentiation of functions can be considered as a mapping in the space of functions. In Mathematica, the differentiation operator is `Derivative`$[1][f]$ or in short form `f'`. If the function `f` is defined, then its derivative can be got by `f'`.

■ *In[1]* := `f`$[x_]$:= `Sin`$[x]$ `Cos`$[x]$ With

In[2] := `f'` follows *Out[2]* = `Cos`$[\#1]^2$ − `Sin`$[\#1]^2$ & ,

hence `f'` is represented as a pure function and it evaluates to

In[3] := % $[x]$ $\longrightarrow$ *Out[3]* = `Cos`$[x]^2$ − `Sin`$[x]^2$

3. Nest The command `Nest`$[f,\ x,\ n]$ means that the function `f` nested n times into itself should be applied on x. The result is `f[f[`$\ldots$`f`$[x]]\ldots]$.

4. NestList By `NestList`$[f,\ x,\ n]$ a list $\{x,$ `f`$[x],$ `f[f`$[x]],\ \ldots\}$ will be shown, where finally f is nested n times. `FoldList`$[f,\ x,\ list]$ iterates a two-variable function.

5. FixedPoint For `FixedPoint`$[f,\ x]$, the function is applied repeatedly until the result does not change.

6. FixedPointList The functional operation `FixedPointList`$[f,\ x]$ shows the continued list with the results after `f` is applied, until the value no longer changes.

■ As an example for this type of functional operation the `NestList` operation will be used for the approximation of a root of an equation $f(x) = 0$ with Newton's method (see 19.1.1.2, p. 950). A root of the equation $x \cos x = \sin x$ is needed in the neighborhood of $3\pi/2$:

In[1] := `f`$[x_]$:= x − `Tan`$[x]$ *In[2]* := `f'`$[x]$ $\longrightarrow$ *Out[2]* = 1 − `Sec`$[x]^2$

In[3] = `g`$[x_]$:= x − `f`$[x]$/`f'`$[x]$

In[4] := `NestList`$[g,\ 4.6,\ 4]$ $\longrightarrow$ *Out[4]* = {4.6, 4.54573, 4.50615, 4.49417, 4.49341}

In[5] := `FixedPoint`$[g,\ 4.6]$ $\longrightarrow$ *Out[5]* = 4.49341

A higher precision of the result can also be achieved.

7. Apply Let f be a function which is considered in connection with a list $\{a,\ b,\ c,\ \ldots\}$. Then

`Apply`$[f,\ \{a,\ b,\ c,\ \ldots\}]$ $\quad f[a,\ b,\ c,\ldots]$ (20.22)

■ *In[1]* := `Apply[Plus,` $\{u,\ v,\ w\}]$ $\longrightarrow$ *Out[1]* = $u + v + w$

In[2] := `Apply[List,` $a + b + c]$ $\longrightarrow$ *Out[2]* = $\{a,\ b,\ c\}$

Here, the general scheme of how Mathematica handles expressions of expressions can be easily recognized. The `FullForm` of the last operation is:

In[3] := `FullForm[Apply[List, Plus`$[a,\ b,\ c]]]$ $\longrightarrow$ *Out[3]* = `List`$[a,\ b,\ c]$

The functional operation `Apply` obviously replaces the `Head` of the considered expression `Plus` by the required `List`.

8. Map With a defined function f the operation `Map` gives:

`Map`$[f,\ \{a,\ b,\ c,\ldots\}]$ $\longrightarrow$ $\{f[a],\ f[b],\ f[c],\ldots\}$ (20.23)

`Map` generates a list whose elements are the values when f is applied to the original list.

■ Let f be the function $f(x) = x^2$. It is defined by

In[1] := `f`$[x_]$:= x^2 With this f one gets

$\texttt{In[2]} := \texttt{Map}[f,\ \{u,\ v,\ w\}] \longrightarrow \texttt{Out[2]} = \{u^2,\ v^2,\ w^2\}$

Map can be applied for more general expressions:

$\texttt{In[3]} := \texttt{Map}[f,\ \texttt{Plus}[a,\ b,\ c]] \longrightarrow \texttt{Out[3]} = a^2 + b^2 + c^2$

20.2.9 Programming

Mathematica can handle the loop constructions known from other languages for procedural programming. The two basic commands are

$$\texttt{Do}[expr,\ \{i, i1, i2, di\}] \quad \text{and} \tag{20.24a}$$

$$\texttt{While}[test,\ expr] \tag{20.24b}$$

The first command evaluates the expression *expr*, where i runs over the values from $i1$ to $i2$ in steps di. If di is omitted, the step size is one. If $i1$ is also missing, then it starts from 1.
The second command evaluates the expression so far as *test* has the value True.

■ In order to determine an approximate value of e^2, the series expansion of the exponential function is used:

$$\begin{aligned} \texttt{In[1]} := \ & sum = 1.0; \\ & \texttt{Do}[sum = sum + (2\hat{}i/i!), \{i, 1, 10\}]; \\ & sum \\ \texttt{Out[1]} = \ & 7.38899 \end{aligned} \tag{20.25}$$

The Do loop evaluates its argument a previously given number of times, while the While loop evaluates as far as a previously given condition becomes false.

Among other things, Mathematica provides the possibility of defining and using local variables. This can be done by the command

$$\texttt{Module}[\{t1, t2, \ldots\},\ procedure] \tag{20.26}$$

The variables or constants enclosed in the list are locally usable in the module; their values assigned here are not valid outside of this module.

■ **A:** A procedure is to be defined which calculates the sum of the square roots of the integers from 1 to n.

$$\begin{aligned} \texttt{In[1]} := \ & sumq[n_] := \\ & \quad \texttt{Module}[\{sum = 1.\}, \\ & \qquad \texttt{Do}[sum = sum + N[\texttt{Sqrt}[i]],\ \{i, 2, n\}]; \\ & \qquad sum \quad]; \end{aligned} \tag{20.27}$$

The call sumq[30] results in 112.083.

The real power of the programming capabilities of Mathematica is, first of all, the use of functional methods in programming, which are made possible by the operations Nest, NestWhile, Apply, Map, MapThread, Distribute and by some further ones.

■ **B:** Example **A** can be written in a functional manner for the case when an accuracy of ten digits is required:

$sumq[n_] := \texttt{N}[\texttt{Apply}[\texttt{Plus}, \texttt{Table}[\texttt{Sqrt}[i], \{i, 1, n\}]], 10]$,

$sumq[30]$ results in 112.0828452. $\texttt{Total}[\sqrt{\texttt{N}[\texttt{Range}[\texttt{n}], \texttt{10}]}]$ gives the same result without using an index, increasing its value continuously, without needing the variable *sum* and its initial value.

For the details, see [20.16].

20.2.10 Supplement about Syntax, Information, Messages

20.2.10.1 Contexts, Attributes

Mathematica must handle several symbols; among them there are those which are used in further program modules loaded on request. To avoid many-valuedness, the names of symbols in Mathematica consist of two parts, the context and the short name.

Short names mean here the names (see 20.2, p. 1024) of heads and elements of the expressions. In addition, in order to name a symbol Mathematica needs the determination of the program part to which the symbol belongs. This is given by the *context*, which holds the name of the corresponding program part. The complete name of a symbol consists of the context and the short name, which are connected by the ' sign.

When Mathematica starts, then there are always two contexts present: *System'* and *Global'*. Information about other available program modules can be obtained by the command `Contexts[]`.

All built-in functions of Mathematica belong to the context *System'*, while the functions defined by the user belong to the context *Global'*.

If a context is actual, thus, the corresponding program part is loaded, then the symbols can be referred to by their short names.
For the input of a further Mathematica program module by << `NamePackage`, the corresponding context is opened and introduced into the previous list. It can happen that a symbol has already been introduced with a certain name before this module is loaded, and in this newly opened context the same name occurs with another definition. In this case Mathematica gives a warning to the user. Then the previously defined name can be erased by the command `Remove[Global‘name]`, or the *complete* name for the newly loaded symbol can be applied.

Besides the properties that the symbols have per definition, it is possible to assign to them some other general properties, called *attributes*, like `Orderless`, i.e., unordered or commutative, `Protected`, i.e., values cannot be changed, or `Locked`, i.e, attributes cannot be changed, etc. Informations about the already existing attributes of the considered object can be obtained by `Attributes`[f].

Some symbols can be protected by `Protect[somesymbol]`; then no other definition can be introduced for this symbol. This attribute can be erased with the command `Unprotect`.

20.2.10.2 Information

Information can be obtained about the fundamental properties of objects by the commands

?*symbol*	information about the object given by the name *symbol*,
??*symbol*	detailed information about the object,
?*B*∗	information about all Mathematica objects, whose names begin with B.

It is also possible to get information about special operators, e.g., by ? := about the SetDelay operator.

However, the most useful possibility is to put the cursor anywhere in the cell containing the symbol of the object in question and press the key F1.

20.2.10.3 Messages

Mathematica has a message system which can be activated and used for different reasons. The messages are generated and shown during the calculations. Their presentation has a uniform form: *symbol* :: *tag*, providing the possibility to refer to them later. (Such messages can also be created by the user.) Consider the following examples as illustrations.

■ **A:** *In[1]* := `f`[x_] := $1/x$; *In[2]* := `f[0]`

`Power: :infy:Infinite expression` $\frac{1}{0}$ `encountered.`

Out[2] = `ComplexInfinity`

■ **B:** *In[1]* := `Log[3, 16, 25]`

`Log: :argt:Log called with 3 arguments; 1 or 2 arguments are expected.`

Out[1] =Log[3, 16, 25]

In example **A**, Mathematica warns us that during the evaluation of an expression it got the value ∞. The calculation itself can be performed. In example **B** the call of logarithm contains three arguments, which is not allowed according to the definition. Calculations cannot be performed. Mathematica cannot do anything with the expression. The user can switch off a message with `Off`[s :: tag]. With `On` the message will appear again. `Quiet` switches off all the messages.
With `Messages`[*symbol*] all messages associated to the symbol with the name *symbol* can be recalled.

20.3 Important Applications with Mathematica

This section describes how to handle mathematical problems with computer algebra systems. The choice of the considered problems is organized according to their frequency in practice and also according to the possibilities of solving them with a computer algebra system. Examples will be given for functions, commands, operations and supplementary syntax. When it is important, the corresponding special package is also discussed briefly.

20.3.1 Manipulation of Algebraic Expressions

In practice, further operations must usually be performed with the occurring algebraic expressions (see 1.1.5, p. 10) such as differentiation, integration, series representation, limiting or numerical evaluation, transformations, etc. In general, these expressions are considered over the ring of integers (see 5.3.7, p. 361) or over the field (see 5.3.7.1, **2.**, p. 361) of real numbers. Computer algebra systems can handle, e.g., polynomials also over finite fields or over extension fields (see 5.3.7.1, **3.**, p. 362) of the rational numbers. Interested people should study the special literature. The algebraic operations with polynomials over the field of rational numbers have special importance. Mathematica provides the functions and operations represented in **Table 20.9** for transformation of algebraic expressions. See also the menu item Palettes|Other|Algebraic Manipulation.

Table 20.9 Commands for manipulation of algebraic expressions

`Expand`[p]	expands the powers and products in a polynomial p by multiplication
`Expand`[p, r]	multiplies only the parts in p, which contain r
`PowerExpand`[a]	expands also the powers of products and powers of powers
`Factor`[p]	factorizes a polynomial completely
`Collect`[p, x]	orders the polynomial with respect the powers of x
`Collect`[$p, \{x, y, \ldots\}$]	the same as the previous one, with several variables
`ExpandNumerator`[r]	expands only the numerator of a rational expression
`ExpandDenominator`[r]	expands only the denominator
`ExpandAll`[r]	expands both numerator and denominator completely
`Together`[r]	combines the terms in the expression over a common denominator
`Apart`[r]	represents the expression in partial fractions
`Cancel`[r]	cancels the common factors in the fraction

20.3.1.1 Multiplication of Expressions

The operation of multiplication of algebraic expressions can always be performed. The coefficients can also be undefined expressions.

■ *In[1]* := `Expand`[$(x + y - z)^4$] gives

Out[1] = $x^4 + 4x^3y + 6x^2y^2 + 4xy^3 + y^4 - 4x^3z - 12x^2yz - 12xy^2z - 4y^3z$
$+6x^2z^2 + 12xyz^2 + 6y^2z^2 - 4xz^3 - 4yz^3 + z^4$

Similarly,

In[3] := `Expand`[$(a\,x + b\,y^2)(c\,x^3 - d\,y^2)$]

Out[3] = $a\,c\,x^4 - a\,d\,x\,y^2 + b\,c\,x^3y^2 - b\,d\,y^4$

20.3.1.2 Factorization of Polynomials

Mathematica performs factorization over the integer or rational numbers if it is possible. Otherwise the original expression is returned.

■ *In[1]* := $p = x^6 + 7x^5 + 12x^4 + 6x^3 - 25x^2 - 30x - 25$;

In[2] := `Factor`$[p]$, gives

Out[2] = $(5 + x)\ (1 + x + x^2)\ (-5 + x^2 + x^3)$

Mathematica decomposes the polynomial into three factors which are irreducible over the rational numbers.

If a polynomial can be completely decomposed over the complex rational numbers, then this can be obtained by the option `GaussianIntegers`.

■ *In[1]* := `Factor`$[x^2 - 2x + 5]$ $\longrightarrow$ *Out[1]* = $5 - 2x + x^2$, but

In[2] := `FactorGaussianIntegers->` `True`]

Out[2] = $((-1 - 2\mathrm{I}) + x)((-1 + 2\mathrm{I}) + x)$

20.3.1.3 Operations with Polynomials

Table 20.10 contains a collection of operations by which polynomials can be algebraically manipulated over the field of rational numbers.

Table 20.10 Algebraic polynomial operations

`PolynomialGCD`$[p1, p2]$	determines the greatest common divisor of $p1$ and $p2$
`PolynomialLCM`$[p1, p2]$	determines the least common multiple of $p1$ and $p2$
`PolynomialQuotient`$[p1, p2, x]$	divides $p1$ (as a function of x) by $p2$, the residue is omitted
`PolynomialRemainder`$[p1, p2, x]$	determines the residue on dividing $p1$ by $p2$
`MonomialList`$[p]$	gives the list of all monomials in the polynomial p

■ Two polynomials are defined:

In[1] := $p = x^6 + 7x^5 + 12x^4 + 6x^3 - 25x^2 - 30x - 25$;
$q = x^4 + x^3 - 6x^2 - 7x - 7$;

With these polynomials the following operations are performed:

In[2] := `PolynomialGCD`$[p, q]$ $\longrightarrow$ *Out[2]* = $1 + x + x^2$

In[3] := `PolynomialLCM`$[p, q]$`//Factor`

Out[3] = $(5 + x)(-7 + x^2)(1 + x + x^2)(-5 + x^2 + x^3)$

In[4] := `PolynomialQuotient`$[p, q, x]$ $\longrightarrow$ *Out[4]* = $12 + 6x + x^2$

In[5] := `PolynomialRemainder`$[p, q, x]$ $\longrightarrow$ *Out[5]* = $59 + 96x + 96x^2 + 37x^3$

With the two last results one gets

$$\frac{x^6 + 7x^5 + 12x^4 + 6x^3 - 25x^2 - 30x - 25}{x^4 + x^3 - 6x^2 - 7x - 7} = x^2 + 6x + 12 + \frac{37x^3 + 96x^2 + 96x + 59}{x^4 + x^3 - 6x^2 - 7x - 7}.$$

20.3.1.4 Partial Fraction Decomposition

Mathematica can decompose a fraction of two polynomials into partial fractions, of course, over the field of rational numbers. The degree of the numerator of any part is always less than the degree of the denominator.

■ Using the polynomials p and q from the previous example one gets

In[1] := `Apart`$[q/p]$ $\longrightarrow$ *Out[1]* = $-\dfrac{6}{35\ (5 + x)} + \dfrac{-55 + 11\,x + 6\,x^2}{35\ (-5 + x^2 + x^3)}$

20.3.1.5 Manipulation of Non-Polynomial Expressions

Complicated expressions, not necessarily polynomials, can often be simplified by the help of the command `Simplify`. Mathematica will always try to manipulate algebraic expressions, independently of the nature of the symbolic quantities. Here, certain built-in knowledge is used. Mathematica knows the rules of powers (see 1.1.4.1, p. 7):

$$\textit{In[1]} := \texttt{Simplify}[a^n/a^m)] \longrightarrow \textit{Out[1]} = a^{-m+n} \qquad (20.28)$$

With the option `Trig −> True`, the commands `Expand` and `Factor` can express powers of trigonometric functions by trigonometric functions with multiple arguments, and conversely. Alternatively, `TrigExpand`, `TrigFactor`, `TrigFactorList`, `TrigReduce`, `ExpToTrig`, `TrigToExp` can be applied.

■ *In[1]* := `TrigExpand[Sin[2x]Cos[2y]]`

Out[1] = `2Cos[x]Cos[y]`^{2}`Sin[x]-2Cos[x]Sin[x]Sin[y]`^{2}

In[2] := `Factor[Sin[4x],Trig−> True]` − `8Cos[x]`^{3}`Sin[x]` + `4Cos[x] Sin[x])`

Out[2] = 0

In[3] := `Factor[Cos[5x],Trig−> True]`

Out[3] = `Cos[x]` (1 − 2 `Cos[2x]` + 2 `Cos[4x]`) .

Remark: The command `ComplexExpand[expr]` assumes a real variable `expr`, while in the command `ComplexExpand`$[expr, \{x1, x2, \ldots\}]$ the variables xi are supposed to be complex.

■ *In[1]* := `ComplexExpand[Sin[2 x],{x}]`
Out[1] = `Cosh[2 Im[x]] Sin[2 Re[x]]+I Cos[2 Re[x]] Sinh[2 Im[x]]`

20.3.2 Solution of Equations and Systems of Equations

Computer algebra systems know procedures to solve equations and systems of equations. If the equation can be solved explicitly in the domain of algebraic numbers, then the solution will be represented with the help of radicals. If it is not possible to give the solution in closed form, then at least numerical solutions can be found with a given accuracy. In the following some basic commands will be introduced. The solution of systems of linear equations (see 4.5.2, p. 308) is discussed here in a special section (see 20.3.2.4, p. 1040).

20.3.2.1 Equations as Logical Expressions

Mathematica allows the manipulation and solution of equations within a wide range. In Mathematica, an equation is considered as a logical expression. If one writes

$$\textit{In[1]} := g = x^2 + 2x - 9 == 0, \qquad (20.29a)$$

then Mathematica considers it as a definition of a function with Boolean values. Giving the input

$$\textit{In[2]} := \%/.\ x{-}{>}\ 2, \ \text{yields } \textit{Out[2]} = \texttt{False}, \qquad (20.29b)$$

since with this value of x the left-hand side and right-hand side are not equal.
The command `Roots`$[g, x]$ transforms the above identity into a form which contains x explicitly. Mathematica represents the result with the help of the logical OR in the form of a logical statement:

$$\textit{In[3]} := \texttt{Roots}[g, x] \longrightarrow \textit{Out[3]} = x == -1 - \sqrt{10} || x == -1 + \sqrt{10} \qquad (20.29c)$$

In this sense, logical operations can be performed with equations.
With the operation `ToRules`, the last logical type equations can be transformed as follows:

In[4] := {`ToRules`[%]}

$$\textit{Out[4]} = \{\{x {-}{>}\ -1 - \sqrt{10}\}, \{x {-}{>} -1 + \sqrt{10}\}\} \qquad (20.29d)$$

20.3.2.2 Solution of Polynomial Equations

Mathematica provides the command `Solve` to solve equations. In a certain sense, `Solve` perform the operations `Roots` and `ToRules` after each other.

Mathematica solves polynomial equations in symbolic form up to fourth degree, since for these equations solutions can be given in the form of algebraic expressions. However, if equations of higher degree can be transformed into a simpler form by algebraic transformations, such as factorization, then Mathematica provides symbolic solutions. In these cases, `Solve` tries to apply the built-in operations `Expand` and `Decompose`.

In Mathematica numerical solutions are also available.

■ The general solution of an equation of third degree:

$$In[1] := \texttt{Solve}[x^3 + a\,x^2 + b\,x + c == 0, x]$$

Mathematica gives

$$Out[1] = \{\{x\text{->} -\frac{a}{3} - \frac{2^{1/3}\,(-a^2+3\,b)}{3\left(-2\,a^3+9\,a\,b-27\,c+3^{3/2}\sqrt{-(a^2\,b^2)+4\,b^3+4\,a^3\,c-18\,a\,b\,c+27\,c^2}\right)^{1/3}} + \frac{\left(-2\,a^3+9\,a\,b-27\,c+3^{3/2}\sqrt{-(a^2\,b^2)+4\,b^3+4\,a^3\,c-18\,a\,b\,c+27\,c^2}\right)^{1/3}}{3 2^{1/3}}\}, \ldots\}$$

The solution list here shows only the first term explicitly because of the length of their terms. If an equation with given coefficients a, b, c has to be solved, then it is better to handle the equation itself with `Solve` than to substitute a, b, c into the solution formula.

■ **A:** For the cubic equation (see 1.6.2.3, p. 40) $x^3 + 6x + 2 = 0$ one gets:

$$In[1] := \texttt{Solve}[x^3 + 6\,x + 2 == 0, x]$$

$$Out[1] = \{\{x\text{->}\, 2^{1/3} - 2^{2/3}\}, \{x\text{->}\, \frac{1-\texttt{I}\sqrt{3}}{2^{1/3}} - \frac{1+\texttt{I}\sqrt{3}}{2^{2/3}}\}, \{x\text{->} -\frac{1-\texttt{I}\sqrt{3}}{2^{2/3}} + \frac{1+\texttt{I}\sqrt{3}}{2^{1/3}}\}\}$$

■ **B:** Solution of an equation of sixth degree:

$$In[2] := \texttt{Solve}[x^6 - 6x^5 + 6x^4 - 4x^3 + 65x^2 - 38x - 120 == 0, x]$$
$$Out[2] = \{\{x\text{->} -1\}, \{x\text{->} -1 - 2\,\texttt{I}\}, \{x\text{->} -1 + 2\,\texttt{I}\}, \{x\text{->}\, 2\}, \{x\text{->}\, 3\}, \{x\text{->}\, 4\}\}$$

Mathematica succeeded in factorizing the equation in **B** with internal tools; then it is solved without difficulty.

If numerical solutions are required, then the command `NSolve` can be used.

■ The following equation is solved by `NSolve`:

$$In[3] := \texttt{NSolve}[x^6 - 4x^5 + 6x^4 - 5x^3 + 3x^2 - 4x + 2 == 0, x]$$
$$Out[3] = \{\{x\text{->} -0.379567 - 0.76948\,\texttt{I}\}, \{x\text{->} -0.379567 + 0.76948\,\texttt{I}\}, \{x\text{->}\, 0.641445\}, \{x\text{->}\, 1. - 1.\,\texttt{I}\}, \{x\text{->}\, 1. + 1.\,\texttt{I}\}, \{x\text{->}\, 2.11769\}\}$$

20.3.2.3 Solution of Transcendental Equations

Mathematica can solve transcendental equations, as well. In general, this is not possible symbolically, and these equations often have infinitely many solutions. In these cases, an estimate of the domain should be given, where Mathematica has to find the solutions. This is possible with the command $\texttt{FindRoot}[g, \{x, x_s\}]$, where x_s is the initial value for the search of the root.

■ $In[1] := \texttt{FindRoot}[x + \texttt{ArcCoth}[x] - 4 == 0, \{x, 1.1\}]$

$Out[1] = \{x\text{->}\, 1.00502\}$ and

$\textit{In[2]} := \texttt{FindRoot}[x + \texttt{ArcCoth}[x] - 4 == 0, \{x, 5\}] \longrightarrow \textit{Out[2]} = \{x{-}{>}\ 3.72478\}$

20.3.2.4 Solution of Systems of Equations

Mathematica can solve simultaneous equations. The operations, built-in for this purpose, are represented in **Table 20.11**, and they present the symbolical solutions, not the numerical ones.
Similarly to the case of one unknown, the command `NSolve` gives the numerical solution(s). The solution of systems of linear equations is discussed in 20.3.3, p. 1040.

Table 20.11 Operations to solve systems of equations

$\texttt{Solve}[\{l_1 == r_1, l_2 == r_2, \ldots\}, vars]$	solves the given system of equations with respect to *vars*
$\texttt{Eliminate}[\{l_1 == r_1, \ldots\}, vars]$	eliminates *vars* from the system of equations
$\texttt{Reduce}[\{l_1 == r_1, \ldots\}, vars]$	simplifies the system of equations and gives the possible solutions
$\texttt{FindInstance}[expr, vars]$	finds an instance of *vars* that make *expr* true

20.3.3 Linear Systems of Equations and Eigenvalue Problems

In 20.2.4, p. 1027, the notion of matrix and several operations with matrices were defined on the basis of lists. Mathematica applies these notions in the theory of systems of linear equations. In the following command m, n denote given integers, not variables.

$$P = \texttt{Array}[p, \{m, n\}] \tag{20.30}$$

defines a matrix of type (m, n) with elements $p_{ij} = \texttt{p}[[i, j]]$. Furthermore

$$X = \texttt{Array}[x, \{n\}] \text{ und } B = \texttt{Array}[b, \{m\}] \tag{20.31}$$

are n- or m-dimensional vectors. With these definitions the general system of linear homogeneous or inhomogeneous equations can be written in the form (see 4.5.2, p. 308)

$$P \,.\, X == B \qquad P \,.\, X == 0 \quad \text{or} \quad \texttt{Thread}[\texttt{P . X == B}] \qquad \texttt{Thread}[\texttt{P . X == 0}] \tag{20.32}$$

1. Special Case $n = m$, $\det P \neq 0$

In the special case $n = m$, $\det P \neq 0$, the system of inhomogeneous equations has a unique solution, which can be determined directly by

$$X = \texttt{Inverse}[P].\, B \tag{20.33}$$

Mathematica can handle such systems of up to ca. 5000 unknowns in a reasonable time, depending on the computer system. An equivalent, but much faster solution is obtained by $\texttt{LinearSolve}[P, B]$.

2. General Case

With the commands `LinearSolve` and `NullSpace`, all the possible cases can be handled as discussed in 4.5.2, p. 308, i.e., it can be determined first if any solution exists, and if it does, then it is calculated. Now, some of the examples from 4.5.2, p. 308ff. will be discussed.

■ **A:** The example in 4.5.2.1, **2.**, p. 310, is a system of homogeneous equations

$$\begin{aligned} x_1 - x_2 + 5x_3 - x_4 &= 0 \\ x_1 + x_2 - 2x_3 + 3x_4 &= 0 \\ 3x_1 - x_2 + 8x_3 + x_4 &= 0 \\ x_1 + 3x_2 - 9x_3 + 7x_4 &= 0 \end{aligned}$$

which has non-trivial solutions. These solutions are the linear combinations of the basis vectors of the null space of matrix p. It is the subspace of the n-dimensional vector space which is mapped into the zero by the transformation p. A basis for this space can be generated by the command $\texttt{NullSpace}[p]$. With the input

$\textit{In[1]} := p = \{\{1, -1, 5, -1\}, \{1, 1, -2, 3\}, \{3, -1, 8, 1\}, \{1, 3, -9, 7\}\}$

a matrix, whose determinant is actually zero is defined (check it by calculating `Det`$[p]$). Now

In[2] := `NullSpace`$[p]$ and *Out[2]* = $\{\{-1,-2,0,1\},\{-3,7,2,0\}\}$

is displayed, a list of two linearly independent vectors of four-dimensional space, which form a basis for the two-dimensional null-space of matrix p. An arbitrary linear combination of these vectors is also in the null-space, so it is a solution of the system of homogeneous equations. This solution coincides with the solution found in 4.5.2.1, **2.**, p. 310.

■ **B:** Consider the example **A** in 4.5.2.1, **2.**, p. 309,

$$\begin{aligned} x_1 - 2x_2 + 3x_3 - x_4 + 2x_5 &= 2 \\ 3x_1 - x_2 + 5x_3 - 3x_4 - x_5 &= 6 \\ 2x_1 + x_2 + 2x_3 - 2x_4 - 3x_5 &= 8 \end{aligned}$$

with matrix $m1$ of type $(3,5)$, and vector $b1$

In[1] := $m1 = \{\{1,-2,3,-1,2\},\{3,-1,5,-3,-1\},\{2,1,2,-2,-3\}\};$

In[2] := $b1 = \{2,6,8\};$

For the command

In[3] := `LinearSolve`$[m1, b1]$ the response is

`LinearSolve` : : nosol: Linear equation encountered which has no solution.

The input appears as output.

■ **C:** According to example **B** from 4.5.2.1, **1.**, p. 309,

$$\begin{aligned} x_1 - x_2 + 2x_3 &= 1 \\ x_1 - 2x_2 - x_3 &= 2 \\ 3x_1 - x_2 + 5x_3 &= 3 \\ -2x_1 + 2x_2 + 3x_3 &= -4 \end{aligned}$$

the input is

In[1] := $m2 = \{\{1,-1,2\},\{1,-2,-1\},\{3,-1,5\},\{-2,2,3\}\};$

In[2] := $b2 = \{1,2,3,-4\};$

To learn how many equations have independent left-hand sides, one calls

In[3] := `RowReduce`$[m2]; \longrightarrow$ *Out[3]* = $\{\{1,0,0\},\{0,1,0\},\{0,0,1\},\{0,0,0\}\}$

Then the input is

In[4] := `LinearSolve`$[m2, b2]; \longrightarrow$ *Out[4]* = $\{\frac{10}{7}, -\frac{1}{7}, -\frac{2}{7}\}$

The answer is the known solution.

3. Eigenvalues and Eigenvectors

Eigenvalues and eigenvectors of matrices are defined in 4.6, p. 314. Mathematica provides the possibility of determining eigenvalues and eigenvectors by special commands. So, the command `Eigenvalues`$[m]$ produces a list of eigenvalues of a square matrix m, `Eigenvectors`$[m]$ creates a list of the eigenvectors of m, whereas `Eigensystem`$[m]$ gives both. If `N`$[m]$ is substituted instead of m, then one gets the numerical eigenvalues. In general, if the order of the matrix is greater than four ($n > 4$), then no algebraic expression can be obtained, since the characteristic polynomial has degree higher than four. In this case, one should ask for numerical values.

■ *In[1]* := $h =$ `Table`$[1/(i+j-1), \{i,5\}, \{j,5\}]$

This generates a five-dimensional so-called Hilbert matrix.

Out[1] = $\{\{1,\frac{1}{2},\frac{1}{3},\frac{1}{4},\frac{1}{5}\},\{\frac{1}{2},\frac{1}{3},\frac{1}{4},\frac{1}{5},\frac{1}{6}\},\{\frac{1}{3},\frac{1}{4},\frac{1}{5},\frac{1}{6},\frac{1}{7}\},\{\frac{1}{4},\frac{1}{5},\frac{1}{6},\frac{1}{7},\frac{1}{8}\},\{\frac{1}{5},\frac{1}{6},\frac{1}{7},\frac{1}{8},\frac{1}{9}\}\}$

With the command

$\mathit{In[2]} := \texttt{Eigenvalues}[h]$

the answer (which may be not useful) is

$\{\texttt{Root}[-1 + 307505\,\#1 - 1022881200\,\#1^2 + \ldots\}$

But with the command

$\mathit{In[3]} := \texttt{Eigenvalues[N[h]]}$ one gets

$\mathit{Out[3]} = \{1.56705, 0.208534, 0.0114075, 0.000305898, 3.28793 \times 10^{-6}\}$

20.3.4 Differential and Integral Calculus

The notation of the derivative as a functional operator was introduced in 20.2.8, p. 1032. Mathematica provides several possibilities to apply the operations of analysis, e.g., determination of the derivative of arbitrarily high order, of partial derivatives, of the complete differential, determination of indefinite and definite integrals, series expansion of functions, and also solutions of differential equations.

20.3.4.1 Calculation of Derivatives

1. Differentiation Operator

The differentiation operator (see 20.2.8, p. 1032) is `Derivative`. Its complete form is

$$\texttt{Derivative}[n_1, n_2, \ldots] \tag{20.34}$$

The arguments say how many times the function is to be differentiated with respect to the current variables. In this sense, it is an operator of partial differentiation. Mathematica tries to represent the result as a pure function.

2. Differentiation of Functions

The differentiation of a given function can be performed in a simplified manner with the operator `D`. With $\texttt{D}[f[x], x]$, the derivative of the function f at the argument x will be determined.
`D` belongs to a group of differential operations, which are enumerated in **Table 20.12**.

Table 20.12 Operations of differentiation

$\texttt{D}[f[x], \{x, n\}]$	yields the n-th derivative of function $f(x)$ with respect to x
$\texttt{D}[f, \{x_1, n_1\}, \{x_2, n_2\}, \cdots]$	multiple derivatives, n_i-th derivative with respect to x_i $(i = 1, 2, \cdots)$
$\texttt{Dt}[f]$	the complete differential of the function f
$\texttt{Dt}[f, x]$	the complete differential $\frac{df}{dx}$ of the function f
$\texttt{Dt}[f, x_1, x_2, \ldots]$	the complete differential of a function of several variables

■ **A** : $\mathit{In[1]} := \texttt{D}[\texttt{Sqrt}[x^3\ \texttt{Exp}[4x]\ \texttt{Sin}[x]], x]$

$$\mathit{Out[1]} = \frac{\texttt{E}^{4x}\, x^3\, \texttt{Cos}[x] + 3\,\texttt{E}^{4x}\, x^2 \texttt{Sin}[\texttt{x}] + 4\texttt{E}^{4x}\, x^3\, \texttt{Sin}[x]}{2\sqrt{\texttt{E}^{4x}\, x^3\, \texttt{Sin}[x]}}$$

■ **B** : $\mathit{In[1]} := \texttt{D}[(2x+1)^{3x}, x] \longrightarrow \mathit{Out[1]} = (1+2\,x)^{3\,x}\left(\frac{6x}{1+2\,x} + 3\texttt{Log[1+2 x]}\right)$

The command `Dt` results in the complete derivative or complete differential.

■ **C** : $\mathit{In[1]} := \texttt{Dt}[x^3 + y^3] \longrightarrow \mathit{Out[1]} = 3x^2\texttt{Dt}[x] + 3y^2\,\texttt{Dt}[y]$

■ **D** : $\mathit{In[1]} := \texttt{Dt}[x^3 + y^3, x] \longrightarrow \mathit{Out[1]} = 3x^2 + 3y^2\texttt{Dt}[y, x]$

In this last example, Mathematica supposes y to be a function of x, which is not known, so it writes the second part of the derivative in a symbolic way. The preferable forms of writing are: $\mathtt{D}[x[t^3] + y[t]^3, t]$ and $\mathtt{D}[x^3 + y[x]^3, x]$ showing explicitly the independent variables.

If Mathematica finds a symbolic function while calculating a derivative, it leaves it in this general form, and expresses its derivative by f'.

■ **E** : *In[1]* := $\mathtt{D}[x\ \mathtt{f}[x]^3, x] \longrightarrow$ *Out[1]* $= \mathtt{f}[x]^3 + 3x\mathtt{f}[x]^2\mathtt{f'}[x]$

Mathematica knows the rules for differentiation of products and quotients, it knows the chain rule, and it can apply these rules formally:

■ **F** : *In[1]* := $\mathtt{D}[\mathtt{f}[\mathtt{u}[x]],\ x] \longrightarrow$ *Out[1]* $= \mathtt{f}'[\mathtt{u}[x]]\ \mathtt{u}'[x]$

■ **G** : *In[1]* := $\mathtt{D}[\mathtt{u}[x]/\mathtt{v}[x], x] \longrightarrow$ *Out[1]* $= \dfrac{\mathtt{u}'[x]}{\mathtt{v}[x]} - \dfrac{\mathtt{u}[x]\ \mathtt{v}'[x]}{\mathtt{v}[x]^2}$

20.3.4.2 Indefinite Integrals

With the command $\mathtt{Integrate}[f, x]$, Mathematica tries to determine the indefinite integral $\int f(x)\,dx$.

If Mathematica knows the integral, it gives it without the integration constant. Mathematica supposes that every expression not containing the integration variable does not depend on it.
In general, Mathematica finds an indefinite integral, if there exists one which can be expressed in closed form by elementary functions, such as rational functions, exponential and logarithmic functions, trigonometric and their inverse functions, etc. If Mathematica cannot find the integral, then it returns the original input. Mathematica knows some special functions which are defined by non-elementary integrals, such as the elliptic functions, and some others.

To demonstrate the possibilities of Mathematica, some examples will be shown, which are discussed in 8.1, p. 480ff.

1. Integration of Rational Functions

(see also 8.1.3.3, p. 485ff.)

■ **A** : *In[1]* := $\mathtt{Integrate}[(2x+3)/(x^3+x^2-2x), x]$

Out[1] $= \dfrac{5}{3}\mathtt{Log}[-1+x] - \dfrac{3\,\mathtt{Log}[x]}{2} - \dfrac{1}{6}\mathtt{Log}[2+x]$

■ **B** : *In[1]* := $\mathtt{Integrate}[(x^3+1)/(x(x-1)^3), x]$

$$\textit{Out[1]} = -\frac{1}{(-1+x)^2} - \frac{1}{-1+x} + 2\,\mathtt{Log}[-1+x] - \mathtt{Log}[x] \qquad (20.35)$$

On the monitor can be seen in the left corner of the next cell a plus sign. Clicking on it one may choose either the free-from input or the Wolfram-Alpha query. If one types the integral into one of these then there is given the possibility to have a look at all the details of the process of integration.

2. Integration of Trigonometric Functions

(see also 8.1.5, p. 491ff.)

■ **A:** The example **A** in 8.1.5.2, p. 492, with the integral $\int \sin^2 x \cos^5 x\,dx$ is calculated (substitution is done by the program automatically, if needed):

In[1] := $\mathtt{Integrate}[\mathtt{Sin}[x]^2\mathtt{Cos}[x]^5, x]$

Out[1] $= \dfrac{5\,\mathtt{Sin}[x]}{64} - \dfrac{1}{192}\mathtt{Sin}[3\,x] - \dfrac{3}{320}\mathtt{Sin}[5\,x] - \dfrac{1}{448}\mathtt{Sin}[7\,x]$

■ **B:** The example **B** in 8.1.5.2, p. 492, with the integral $\int \frac{\sin x}{\sqrt{\cos x}}\,dx$ is calculated:

$$\textit{In[1]} := \texttt{Integrate}[\texttt{Sin}[x]/\texttt{Sqrt}[\texttt{Cos}[x]], x] \longrightarrow \textit{Out[1]} = -2\sqrt{\texttt{Cos}[x]}$$

Remark: In the case of non-elementary integrals Mathematica may do nothing.

■ $\textit{In[1]} := \int x^x \mathrm{dx} \longrightarrow \textit{Out[1]} = \int \mathrm{x}^{\mathrm{x}} \mathrm{dx}$

20.3.4.3 Definite Integrals and Multiple Integrals

1. Definite Integrals

With the command `Integrate`$[f, \{x, x_a, x_e\}]$, Mathematica can evaluate the definite integral of the function $f(x)$ with a lower limit x_a and upper limit x_e.

■ **A** : $\textit{In[1]} := \texttt{Integrate}[\texttt{Exp}[-x^2], \{x, 0, \texttt{Infinity}\}] \longrightarrow \textit{Out[1]} = \frac{\sqrt{\pi}}{2}$

(see **Table 21.8**, p. 1098, No. 25 for $a = 1$).

■ **B:** If the input is

$\textit{In[1]} := \texttt{Integrate}[\frac{1}{x^2}, \{x, -1, 1\}]$ one gets

$\textit{Out[1]} =$ Integrate::idiv: "Integral of $\frac{1}{x^2}$ does not converge on {-1,1}. "

In the calculation of definite integrals one should be careful. If the properties of the integrand are not known, then it is recommended to ask for a graphical representation of the function in the considered domain before integration.

2. Multiple Integrals

Definite double integrals can be called by the command

$$\texttt{Integrate}[\texttt{f}[x, y], \{x, x_a, x_e\}, \{y, y_a, y_e\}] \qquad (20.36)$$

The evaluation is performed from right to left, so, first the integration is evaluated with respect to y. The limits y_a and y_e can be functions of x, which are substituted into the primitive function. Then the integral is evaluated with respect to x.

■ For the integral A, which calculates the area between a parabola and a line intersecting it twice, in 8.4.1.2, p. 524, one gets

$$\textit{In[1]} := \texttt{Integrate}[x\ y^2, \{x, 0, 2\}, \{y, x^2, 2x\}] \longrightarrow \textit{Out[1]} = \frac{32}{5}.$$

Also in this case, it is important to be careful with the discontinuities of the integrand. The domain of integration can also be specified with inequalities: `Integrate[Boole[`$x^2+y^2 \le 1, \{x, -1, 1\}, \{y, -1, 1\}$`]]` gives π.

20.3.4.4 Solution of Differential Equations

Mathematica can handle ordinary differential equations symbolically if the solution can be given in closed form. In this case, Mathematica gives the solution in general. The commands discussed here are listed in **Table 20.13**.

The solutions (see 9.1, p. 540) are represented as general solutions with the arbitrary constants $C[i]$. Initial values and boundary conditions can be introduced in the part of the list which contains the equation or equations. In this case a special solution is returned. As examples, two differential equations are solved here from 9.1.1.2, p. 542.

■ **A:** The solution of the differential equation $y'(x) - y(x)\tan x = \cos x$ is to be determined.

$$\textit{In[1]} := \texttt{DSolve}[y'[x] - y[x]\ \texttt{Tan}[x] == \texttt{Cos}[x], y, x]$$

Mathematica solves this equation, and gives the solution as a pure function with the integrations constant $C[1]$.

Table 20.13 Commands to solve differential equations

`DSolve`$[deq, y[x], x]$	solves the differential equation for $y[x]$ (if it is possible); $y[x]$ may be given in implicit form
`DSolve`$[deq, y, x]$	gives the solution of the differential equation in the form of a pure function
`DSolve`$[\{deq_1, deq_2, \ldots\}, y, x]$	solves a system of ordinary differential equations

$$\textit{Out[1]} = \left\{\left\{y \;\rightarrow\; \texttt{Function[x, C[1] Sec[x] + Sec[x]} \left(\frac{x}{2} + \frac{1}{4}\texttt{Sin[2 x]}\right)]\right\}\right\}$$

If it is required to get the solution value $y[x]$, then Mathematica gives

$$\textit{In[2]} := y[x]/.\ \%1 \;\longrightarrow\; \textit{Out[2]} = \left\{\mathrm{C}[1]\texttt{Sec}[x] \;+\texttt{Sec}[x] \left(\frac{x}{2} + \frac{1}{4}\texttt{Sin}[2\,x]\right)\right\}$$

One also could make the substitution for other quantities, e.g., for $y'[x]$ or $y[1]$. The advantage of using pure functions is obvious here.

■ **B:** The solution of the differential equation $y'(x)x(x - y(x)) + y^2(x) = 0$ (see 9.1.1.2, **2.**, p. 542) is to be determined.

$$\textit{In[1]} := \texttt{DSolve}[y'[x]\; x(x - y[x]) + y[x]\hat{\ }2 == 0, y[x], x]$$

$$\textit{Out[1]} = \left\{\left\{y[x] \;\longrightarrow\; \texttt{-x ProductLog}[-\frac{E^{-\mathrm{C}[1]}}{x}]\right\}\right\}$$

Here `ProductLog[z]` gives the principal solution for w in $z = we^w$. The solution of this differential equation was given in implicit form (see 9.1.1.2, **2.**, p. 542).

If Mathematica cannot solve a differential equation it returns the input without any comment. In such cases, or also, if the symbolic solution is too complicated, the solutions can be found by numerical solutions (see 19.8.4.2, **5.**, p. 1018). Also in the case of symbolic solutions of differential equations, like in the evaluation of indefinite integrals, the efficiency of Mathematica should not be overestimated. If the result cannot be expressed as an algebraic expression of elementary functions, the only way is to find a numerical solution.

Remark: Mathematica can solve some partial differential equations both symbolically and numerically, as well, even on complicated multidimensional domains.

20.4 Graphics with Mathematica

By providing routines for graphical representation of mathematical relations such as the graphs of functions, space curves, and surfaces in three-dimensional space, modern computer algebra systems provide extensive possibilities for combining and manipulating formulas, especially in analysis, vector calculus, and differential geometry, and they provide immeasurable help in engineering designing. Graphics is a special strength of Mathematica.

20.4.1 Basic Elements of Graphics

Mathematica builds graphical objects from built-in *graphics primitives*. These are objects such as points (`Point`), lines (`Line`) and polygons (`Polygon`) and properties of these objects such as thickness and color.

Mathematica has several options to specify the environment for graphics and how the graphical objects should be represented.

With the command Graphics[*list*], where *list* is a list of graphics primitives, Mathematica is called to generate a graphic from the listed objects. The object list can follow a list of options about the appearance of the representation.
With the input

$$\textit{In[1]} := g = \texttt{Graphics}[\{\texttt{Line}[\{\{0,0\},\{5,5\},\{10,3\}\}],\texttt{Circle}[\{5,5\},4], \qquad (20.37a)$$

$$\texttt{Text}[\texttt{Style}[\text{“Example”},\text{“Helvetica”},\text{Bold},25],\{5,6\}]\},\texttt{AspectRatio}\text{–>}\ \texttt{Automatic}] \qquad (20.37b)$$

a graphic is built from the following elements:

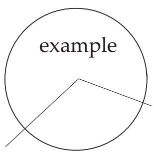


Figure 20.1

a) Broken line of two line segments starting at the point $(0,0)$ through the point $(5,5)$ to the point $(10,3)$.

b) Circle with the center at $(5,5)$ and radius 4.

c) Text with the content “Example”, written in Helvetica font, boldface (the text appears centered with respect to the reference point $(5,6)$).
With the call Show[g], Mathematica displays the figure **(Fig. 20.1)**.
Certain options might be previously specified. Here the option AspectRatio is set to Automatic.

By default Mathematica makes the ratio of the height to the width of the graph 1 : GoldenRatio (see e.g. 3.5.2.3,**3.**, p. 194). It corresponds to a relation between the extension in the x direction to the one in $1 : 1/1.618 = 1 : 0.618$. With this option the circle would be deformed into an ellipse. The value of the option Automatic ensures that the representation is not deformed.

20.4.2 Graphics Primitives

Mathematica provides the two-dimensional graphic objects enumerated in **Table 20.14**.
Besides these objects Mathematica provides further primitives to control the appearance of the representation, the graphics commands. They specify how graphic objects should be represented. The commands are listed in **Table 20.15**.
There is a wide scale of colors to choose from but their definitions are not discussed here.

Table 20.14 Two-dimensional graphic objects

Point[$\{x,\ y\}$]	point at position x, y
Line[$\{\{x_1,y_1\},\{x_2,y_2\},\ldots\}$]	broken line through the given points
Rectangle[$\{x_{lu},y_{lu}\},\{x_{ro},y_{ro}\}$]	shaded rectangle with the given coordinates left-down, right-up
Polygon[$\{\{x_1,y_1\},\{x_2,y_2\},\ldots\}$]	shaded polygon with the given vertices
Circle[$\{x,y\},r$]	circle with radius r around the center x, y
Circle[$\{x,y\},r,\{\alpha_1,\alpha_2\}$]	circular arc with the given angles as limits
Circle[$\{x,y\},\{a,b\}$]	ellipse with half-axes a and b
Circle[$\{x,y\},\{a,b\},\{\alpha_1,\alpha_2\}$]	elliptic arc
Disk[$\{x,y\},r$], Disk[$\{x,y\},\{a,b\}$]	shaded circle or ellipse
Text[$text$, $\{x,y\}$]	writes *text* centered to the point x, y

Table 20.15 Graphics commands

PointSize[a]	a dot is drawn with radius a as a fraction of the total picture
AbsolutePointSize[b]	denotes the absolute radius b of the dot (measured in American pt (0.3515 mm))
Thickness[a]	draws lines with relative thickness a
AbsoluteThickness[b]	draws lines with absolute thickness b (also in pt)
Dashing[$\{a_1,a_2,a_3,\ldots\}$]	draws a line as a sequence of stripes with the given length (in relative measure)
AbsoluteDashing[$\{b_1,b_2,\ldots\}$]	the same as the previous one but in absolute measure
GrayLevel[p]	specifies the level of shade ($p = 0$ is for black, $p = 1$ is for white)

20.4.3 Graphical Options

Mathematica provides several graphical options which have an influence on the appearance of the entire picture. **Table 20.16** gives a selection of the most important commands. For a detailed explanation, see [20.16].

Table 20.16 Some graphical options

`AspectRatio ->` w	sets the ratio w of height and width. `Automatic` determines w from the absolute coordinates; the default setting is $w = 1$: `GoldenRatio`
`Axes -> True`	draws coordinate axes
`Axes -> False`	does not draw coordinate axes
`Axes -> {True, False}`	shows only the x-axis
`Frame -> True`	shows frames
`GridLines -> Automatic`	shows grid lines
`AxesLabel ->` $\{x_{symbol}, y_{symbol}\}$	denotes axes with the given symbols
`Ticks -> Automatic`	denotes scaling marks automatically; with `None` they can be suppressed
`Ticks ->` $\{\{x_1, x_2, \ldots\}, \{y_1, y_2, \ldots\}\}$	scaling marks are placed at the given nodes

20.4.4 Syntax of Graphical Representation

20.4.4.1 Building Graphic Objects

If a graphic object is to be built from primitives, then first a list of the corresponding objects with their global definition should be given in the form

$$\{object_1, object_2, \ldots\}, \qquad (20.38a)$$

where the objects themselves can be lists of graphic objects. Let object1 be, e.g.,

In[1] $:= o1 = \{$`Circle`$[\{5,5\},\{5,3\}],$ `Line`$[\{\{0,5\},\{10,5\}\}]\}$

and corresponding to it

In[2] $:= o2 = \{$`Circle`$[\{5,5\},3]\}$

as in **Fig.20.1**. If a graphic object, e.g., $o2$, is to be provided with certain graphical commands, then it should be written into one list with the corresponding command

In[3] $:= o3 = \{$`Thickness`$[0.01], o2\}$

This command is valid for all objects in the *corresponding* braces, and also for nested ones, but not for the objects outside of the braces of the list.
From the generated objects two different graphic lists are defined:

In[4] $:= g1 =$ `Graphics`$[\{o1, o2\}]$; $g2 =$ `Graphics`$[\{o1, o3\}]$

which differs only in the second object by the thickness of the circle. The call

`Show`$[g1]$ and `Show`$[g2,$ `Axes -> True`$]$ (20.38b)

givs the pictures represented in **Fig. 20.2**.

In the call of the picture in **Fig. 20.2b**, the option `Axes -> True` was activated. This results in the representation of the axes with marks on them chosen by Mathematica and with the corresponding scaling.

20.4.4.2 Graphical Representation of Functions

Mathematica has special commands for the graphical representation of functions. With

`Plot`$[$`f`$[x], \{x, x_{min}, x_{max}\}]$ (20.39)

the function f is represented graphically in the domain between $x = x_{min}$ and $x = x_{max}$. Mathematica produces a function table by internal algorithms and reproduces the graphics following from this table by graphics primitives.

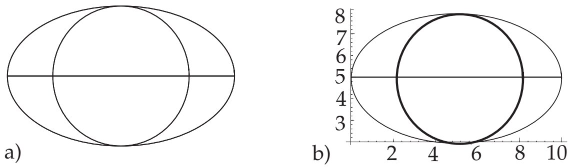


Figure 20.2

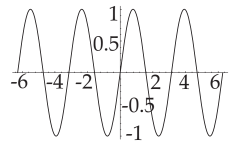


Figure 20.3

■ If the function $x \mapsto \sin 2x$ is to be graphically represented in the domain between -2π and 2π, then the input is

In[1] := Plot[Sin[$2x$], {x, −2Pi, 2Pi}].

Mathematica produces the curve shown in **Fig. 20.3**.

It is obvious that Mathematica uses certain default graphical options in the representation as mentioned in 20.4.1, p. 1045. So, the axes are automatically drawn, they are scaled and denoted by the corresponding x and y values. In this example, the influence of the default AspectRatio can be seen. The ratio of the total width to the total height is 1 : 0.618.

With the command InputForm[%] the whole representation of the graphic objects can be shown. For the previous example one gets:

```
Graphics[{{{}, {}, {Directive[Opacity[1.], RGBColor[0.368417, 0.506779, 0.709798],
AbsoluteThickness[1.6]], Line[{{-6.283185050723043, 2.5645654335783057*^-7},
..., {6.283185050723043, -2.5645654335783057*^-7} }], {DisplayFunction -> Identity,
AspectRatio -> GoldenRatio^(-1), Axes -> {True, True}, AxesLabel -> {None, None},
AxesOrigin -> {0, 0}, DisplayFunction :> Identity,
Frame -> {{False, False}, {False, False}}, FrameLabel -> {{None, None},
{None, None}}, FrameTicks -> {{Automatic, Automatic}, {Automatic, Automatic}},
GridLines -> {None, None}, GridLinesStyle -> Directive[GrayLevel[0.5, 0.4]],
Method -> {"DefaultBoundaryStyle" -> Automatic, "ScalingFunctions" -> None},
PlotRange -> {{-2*Pi, 2*Pi}, {-0.9999996654606427, 0.9999993654113022}},
PlotRangeClipping -> True, PlotRangePadding -> {{Scaled[0.02], Scaled[0.02]},
{Scaled[0.05], Scaled[0.05]}}, Ticks -> {Automatic, Automatic}}]
```

Consequently, the graphic object consists of a few sublists. The first one contains the graphics primitive Line (slightly modified), with which the internal algorithm connects the calculated points of the curve by lines. The second sublist contains the options needed by the given graphic. These are the default options. If the picture is to be altered at certain positions, then the new settings in the Plot command must be set after the main input. With

In[2] := Plot[Sin[$2x$], {x, −2Pi, 2Pi}, AspectRatio−> 1] (20.40)

the representation would be done with equal length of axes x and y.
It is possible to give several options at the same time after each other.With the input

Plot[{$f_1[x]$, $f_2[x]$, . . .}, {x, x_{min}, x_{max}}] (20.41)

several functions are shown in the same graphic. With the command

Show[*plot*, *options*] (20.42)

an earlier picture can be renewed with other options.With

$\texttt{Show}[\texttt{GraphicsArray}[list]]$, (20.43)

(with *list* as lists of graphic objects) pictures can be placed next to each other, under each other, or they can be arranged in matrix form.

20.4.5 Two-Dimensional Curves

A series of curves from the chapter on functions and their representations (see 2.1, p. 48ff.) is shown as examples.

20.4.5.1 Exponential Functions

A family of curves with several exponential functions (see 2.6.1, p. 72) is generated by Mathematica **(Fig. 20.4a)** with the following input:

$\textit{In[1]} := \texttt{f}[x_] := 2\hat{}x; \texttt{g}[x_] := 10\hat{}x;$

$\textit{In[2]} := \texttt{h}[x_] := (1/2)\hat{}x; \texttt{j}[x_] := (1/\texttt{E})\hat{}x; \texttt{k}[x_] := (1/10)\hat{}x;$

These are the definitions of the considered functions. There is no need to define the function e^x, since it is built into Mathematica. In the second step the following graphics are generated:

$\textit{In[3]} := p1 = \texttt{Plot}[\{\texttt{f}[x], \texttt{h}[x]\}, \{x, -4, 4\}, \textit{PlotStyle}{-}{>}\ \texttt{Dashing}[\{0.01, 0.02\}]]$

$\textit{In[4]} := p2 = \texttt{Plot}[\{\texttt{Exp}[x], \texttt{j}[x]\}, \{x, -4, 4\}]$

$\textit{In[5]} := p3 = \texttt{Plot}[\{\texttt{g}[x], \texttt{k}[x]\}, \{x, -4, 4\}, \texttt{PlotStyle}{-}{>}\ \texttt{Dashing}[\{0.005, 0.02, 0.01, 0.02\}]]$

The whole picture **(Fig. 20.4a)** can be obtained by:

$\textit{In[6]} := \texttt{Show}[\{p1, p2, p3\}, \texttt{PlotRange}{-}{>}\ \{0, 18\}, \texttt{AspectRatio}{-}{>}\ 1.2]$

The question of how to write text on the curves is not discussed here. This is possible with the graphics primitive `Text`.

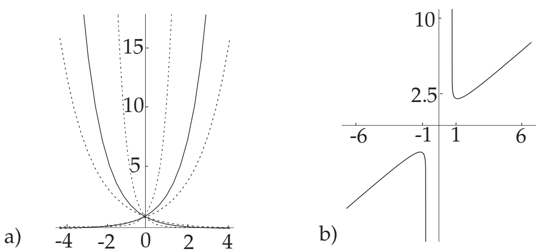


Figure 20.4

20.4.5.2 Function $y = x + \operatorname{Arcoth} x$

Considering the properties of the function $\operatorname{Arcoth} x$ discussed in 2.10, p. 93, the function $y = x + \operatorname{Arcoth} x$ can be graphically represented in the following way:

$\textit{In[1]} := f1 = \texttt{Plot}[x + \texttt{ArcCoth}[x], \{x, 1.000000000005, 7\}]$

$\textit{In[2]} := f2 = \texttt{Plot}[x + \texttt{ArcCoth}[x], \{x, -7, -1.000000000005\}]$

$\textit{In[3]} := \texttt{Show}[\{f1, f2\}, \texttt{PlotRange}{-}{>}\ \{-10, 10\}, \texttt{AspectRatio}{-}{>}\ 1.2, \texttt{Ticks}{-}{>}$

$\{\{\{-6, -6\}, \{-1, -1\}, \{1, 1\}, \{6, 6\}\}, \{\{2.5, 2.5\}, \{10, 10\}\}\}, \texttt{AxesOrigin}- > 0, 0]$

The high precision of the x values in the close neighborhood of 1 and -1 was chosen to get sufficiently large function values for the required domain of y. The result is shown in **Fig. 20.4b**.

20.4.5.3 Bessel Functions (see 9.1.2.6, **2.**, p. 562)

With the calls

In[1] := `bj0 = Plot[{BesselJ[0, z], BesselJ[2, z], BesselJ[4, z]}], {z, 0, 10}, PlotLabel–>`
`TraditionalForm[{BesselJ[0, z], BesselJ[2, z], BesselJ[4, z]}]` (20.44a)

In[2] := `bj1 = Plot[{BesselJ[1, z], BesselJ[3, z], BesselJ[5, z]}, {z, 0, 10}, PlotLabel–>`
`TraditionalForm[{BesselJ[1, z], BesselJ[3, z], BesselJ[5, z]}]]` (20.44b)

the graphics of the Bessel function $J_n(z)$ for $n = 0, 2, 4$ and $n = 1, 3, 5$ are generated, which are then represented by the call

In[3] := `GraphicsRow[{bj0, bj1}]]`

next to each other in **Fig. 20.5**.

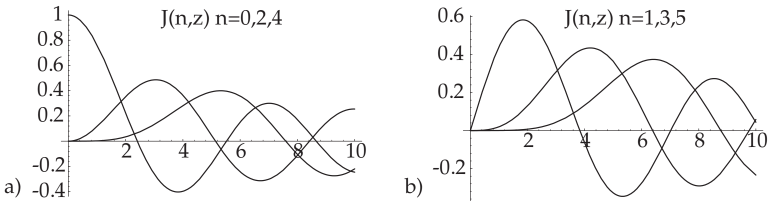


Figure 20.5

20.4.6 Parametric Representation of Curves

Mathematica has a special graphics command, with which curves given in parametric form can be graphically represented. This command is:

`ParametricPlot[{`$f_x(t), f_y(t)$`}, {`t, t_1, t_2`}]`. (20.45)

It provides the possibility of showing several curves in one graphic. A list of several curves must be given in the command. With the option `AspectRatio–> Automatic`, Mathematica shows the curves in their natural forms.

The parametric curves in **Fig. 20.6** are the Archimedean spiral (see 2.14.1, p. 105) and the logarithmic spiral (see 2.14.3, p. 106). They are represented with the input

In[1] := `ParametricPlot[{t Cos[t], t Sin[t]}, {t, 0, 3Pi}, AspectRatio–> Automatic]`

and

In[2] := `ParametricPlot[{Exp[0.1t] Cos[t], Exp[0.1t] Sin[t]}, {t, 0, 3Pi},`
`AspectRatio–> Automatic]`

With

In[3] := `ParametricPlot[{t − 2 Sin[t], 1 − 2 Cos[t]}, {t, −Pi, 11Pi}, AspectRatio–> 0.3]`

a trochoid (see 2.13.2, p. 102) is generated **(Fig. 20.7)**.

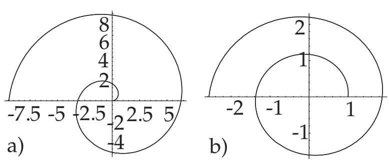


Figure 20.6

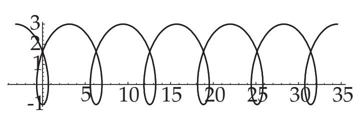


Figure 20.7

20.4.7 Representation of Surfaces and Space Curves

Mathematica provides the possibility of representing three-dimensional graphics primitives.
Similarly to the two-dimensional case, three-dimensional graphics can be generated by applying different options. The objects can be represented and observed from different viewpoints and from different perspectives. Also the representation of curved surfaces in three-dimensional space, i.e., the graphical representation of functions of two variables, is possible. Furthermore it is possible to represent curves in three-dimensional space, e.g., if they are given in parametric form. For a detailed description of three-dimensional graphics primitives see [20.5], [20.16]. The introduction of these representations is similar to the two-dimensional case.

20.4.7.1 Graphical Representation of Surfaces

The command `Plot3D` in its basic form requires the definition of a function of two variables and the domain of these two variables:

$$\texttt{Plot3D}[\texttt{f}[x,y],\{x,x_a,x_e\},\{y,y_a,y_e\}] \tag{20.46}$$

All options have the default setting.

■ For the function $z = x^2 + y^2$, with the input

$\textit{In[1]} := \texttt{Plot3D}[x^2+y^2,\{x,-5,5\},\{y,-5,5\},\texttt{PlotRange}\text{->}\{0,25\}]$

we get **Fig. 20.8a**, while **Fig. 20.8b** is generated by the command

$\textit{In[2]} := \texttt{Plot3D}[(1-\texttt{Sin}[x])\;(2-\texttt{Cos}[2\;y]),\{x,-2,2\},\{y,-2,2\}]$

For the paraboloid, the option `PlotRange` is given with the required z values, because the solid is cut at $z = 25$.

20.4.7.2 Options for 3D Graphics

The number of options for 3D graphics is large. In **Table 20.17**, only a few are enumerated, where options known from 2D graphics are not included. They can be applied in a similar sense. The option `ViewPoint` has special importance, by which very different observational perspectives can be chosen.

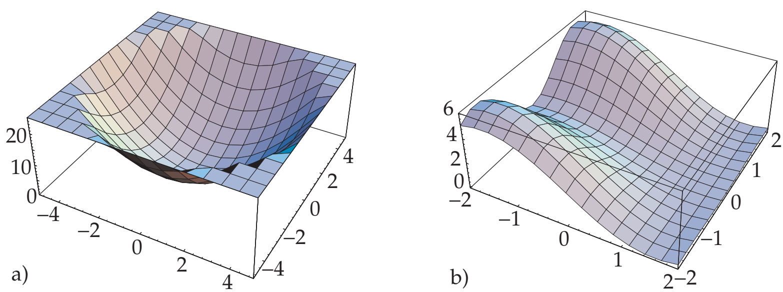


Figure 20.8

20.4.7.3 Three-Dimensional Objects in Parametric Representation

Similarly to 2D graphics, three-dimensional objects given in parametric representation can also be represented. With

$$\texttt{ParametricPlot3D}[\{f_x[t,u],f_y[t,u],f_z[t,u]\},\{t,t_a,t_e\},\{u,u_a,u_e\}] \tag{20.47}$$

a parametrically given surface is represented, with

$$\texttt{ParametricPlot3D}[\{f_x[t],f_y[t],f_z[t]\},\{t,t_a,t_e\}] \tag{20.48}$$

a three-dimensional curve is generated parametrically.

Table 20.17 Options for 3D graphics

`Boxed`	default setting is `True`; it draws a three-dimensional frame around the surface
`HiddenSurface`	sets the non-transparency of the surface; default setting is `True`
`ViewPoint`	specifies the point (x, y, z) in space, from where the surface is observed. Default values are $\{1.3, -2.4, 2\}$
`Shading`	default setting is `True`; the surface is shaded; `False` yields white surfaces
`PlotRange`	$\{z_a, z_e\}$, $\{\{x_a, x_e\}, \{y_a, y_e\}, \{z_a, z_e\}\}$ can be chosen for the values `All`. Default is `Automatic`

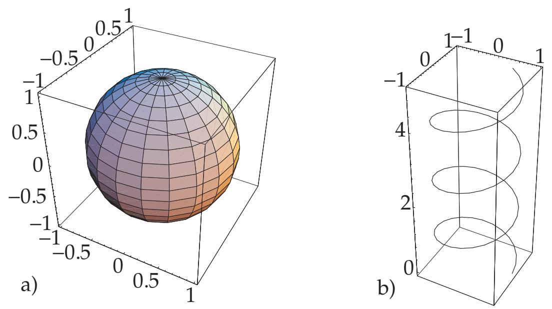


Figure 20.9

■ The objects in **Fig. 20.9a** and **Fig. 20.9b** are represented with the commands

$$\textit{In[3]} := \texttt{ParametricPlot3D}[\{\texttt{Cos}[t]\ \texttt{Cos}[u], \texttt{Sin}[t]\ \texttt{Cos}[u], \texttt{Sin}[u]\}, \{t, 0, 2\texttt{Pi}\}\ \{u, -\texttt{Pi}/2, \texttt{Pi}/2\}] \tag{20.49a}$$

$$\textit{In[4]} := \texttt{ParametricPlot3D}[\{\texttt{Cos}[t], \texttt{Sin}[t], t/4\}, \{t, 0, 20\}] \tag{20.49b}$$

Mathematica provides further commands by which density, and contour diagrams, bar charts and sector diagrams, and also a combination of different types of diagrams, can be generated.

■ The representation of the Lorenz attractor (see 17.2.4.3, p. 887) can easily be generated by Mathematica.

There is a series of recent developments most of which are not to be shown in a book. One can easily build a GUI (graphical user interface) to utilize interactive properties of the program. Most of the calculations are parallelized automatically, but functions such as `Parallelize` and `ParallelMap` provides the user to create his/her own parallel programs. The extremely fast graphic cards can be programmed at a very high level (as opposed to other languages) using such functions as `CUDALink`, `OpenCLFunctionLoad` etc. An extremely useful example of dynamic interactivity tool is `Manipulate` which in the simplest case shows you the parameter dependence of a family of curves. Working in the cloud or using the computer Raspberry Pi (which comes a free Mathematica license) should also not be unmentioned.

21 Tables

21.1 Frequently Used Mathematical Constants

π	3,141592654...	Ludolf constant (π)	$1°/_{\circ}$	0,01	percent
e	2,718281828...	Euler constant (e)	$1°/_{\circ\circ}$	0,001	per mil
C	0,577215665...	Euler constant (C)	$\sqrt{2}$	1,414 2136...	
$\lg e = M$	0,434294482...	$\ln 10 = M^{-1} = 2{,}302585093\ldots$	$\sqrt{3}$	1,732 0508...	
$\lg 2$	0,301 030...	$\ln 2 = 0{,}693\,1472\ldots$	$\sqrt{10}$	3,162 2777...	

21.2 Important Natural Constants

This table contains values of constants, recommended in [21.19], [21.20], [21.21]. In parenthesis is given the standard uncertainty of the last two digits. The note **(fixed)** indicates that this value is fixed by definition.

Fundamental constants		
Avogadro constant	N_A	$= 6{,}022\,141\,29(27)\cdot 10^{23}/\mathrm{mol}$
velocity of light in vacuum	c_0	$= 299\,792\,458\mathrm{m/s}$ **(fixed)**
gravitation constant	G	$= 6{,}673\,84\,(80)\cdot 10^{-11}\mathrm{m}^3/(\mathrm{kg\,s}^2)$
fundamental electric charge	e	$= 1{,}602\,176\,565(35)\cdot 10^{-19}\ \mathrm{C}$
fine structure constant	α	$= \mu_0 c_0 e^2/(2h) = 7{,}297\,352\,5698(24)\cdot 10^{-3}$
Sommerfeld constant	α^{-1}	$= 137{,}035\,999\,074(44)$
Planck constant	h	$= 6{,}626\,069\,57(29)\cdot 10^{-34}\mathrm{Js} = 4{,}135\,667\,516(91)\cdot 10^{-15}\mathrm{eVs}$
Planck quantum $h/(2\pi)$	$\hbar$	$= 1{,}054\,571\,726(47)\cdot 10^{-34}\mathrm{Js}$
		$= 6{,}582\,119\,28(15)\cdot 10^{-16}\mathrm{eVs}$

Electromagnetic constants		
spec. fund. electric charge	$-e/m_\mathrm{e}$	$= -1{,}758\,820\,088(39)\cdot 10^{11}\ \mathrm{C\,kg}^{-1}$
permeability of free space	μ_0	$= 4\pi\cdot 10^{-7}\ \mathrm{N/A}^2 = 12{,}566\,370\,614\cdot 10^{-7}\ \mathrm{Vs/Am}$ **(fixed)**
permittivity of vacuum	ε_0	$= 1/(\mu_0 c_0^2) = 8{,}854\,187\,817\cdot 10^{-12}\ \mathrm{As/Vm}$ **(fixed)**
quantum of magnetic flux	Φ_0	$= h/(2e) = 2{,}067\,833\,758(46)\cdot 10^{-15}\ \mathrm{Wb}$
Josephson constant	K_J	$= 2e/h = 483\,597{,}870(11)\cdot 10^{9}\ \mathrm{Hz/V}$
v. Klitzing constant	R_K	$= h/e^2 = 25\,812{,}807\,4434(84)\ \Omega$
quantum of conductance	G_0	$= 2e/h = 7{,}748\,091\,7346(25)\cdot 10^{-5}\ \mathrm{S}$
character. impedance (vacuum)	Z_0	$= 376{,}730\,313\,461\ \ \Omega$ **(fixed)**
Faraday constant	F	$= eN_\mathrm{A} = 96\,485{,}3365(21)\ \mathrm{As/mol}$

Constants in physical chemistry, thermodynamics, mechanics		
Boltzmann constant	k	$= R_0/N_\mathrm{A} = 1{,}380\,6488(13)\cdot 10^{-23}\mathrm{J/K}$
		$= 8{,}617\,3324(78)\cdot 10^{-5}\mathrm{eV/K}$
universal gas constant, molar	R_0	$= N_\mathrm{A}k = 8{,}314\,4621(75)\ \mathrm{J/(mol\,K)}$
molar volume of inert gas (T_0 =273,15 K, p_0 =100 kPa	V_{m_0}	$= R_0T_0/p_0 = 22{,}710\,953(21)\cdot 10^{-3}\ \mathrm{m}^3/\mathrm{mol}$
molar volume of inert gas (T_0 =273,15 K, p_1 =101,325 kPa	V_{m_1}	$= R_0T_0/p_0 = 22{,}413\,968(20)\cdot 10^{-3}\ \mathrm{m}^3/\mathrm{mol}$
Loschmidt constant (T_0, p_0)	n_{00}	$= N_\mathrm{A}/V_{\mathrm{m}_0} = 2{,}651\,6462(24)\cdot 10^{25}/\mathrm{m}^3$
Loschmidt constant (T_0, p_1)	n_{01}	$= N_\mathrm{A}/V_{\mathrm{m}_1} = 2{,}686\,7805(24)\cdot 10^{25}/\mathrm{m}^3$
standard acceleration of gravity (earth, 45° geographic latitude, sea level)	g_n	$= 9{,}806\,65\ \mathrm{m\,s}^{-2}$ **(fixed)**

I.N. Bronshtein et al., *Handbook of Mathematics*,
DOI 10.1007/978-3-662-46221-8_21

Atomic electron shell and atomic nucleus		
atomic mass unit u	m_u	$= (10^{-3}\mathrm{kg/mol})/N_\mathrm{A} = \frac{1}{12} m_\mathrm{atom}(^{12}\mathrm{C})$
		$= 1{,}660\,538\,921(73) \cdot 10^{-27}$ kg
quantum of circulation (electron)	s	$= h/(2m_\mathrm{e}) = 3{,}636\,947\,5520(24) \cdot 10^{-4}\ \mathrm{m^2/s}$
Bohr radius	a_0	$= \hbar^2/(E_0(e)e^2) = r_e/\alpha^2 = 0{,}529\,177\,210\,92(17) \cdot 10^{-10}$ m
classical electron radius	r_e	$= \alpha^2 a_0 = 2{,}817\,940\,3267(27) \cdot 10^{-15}$ m
Thomson cross-section	σ_0	$= 8\pi r_\mathrm{e}^2/3 = 0{,}665\,245\,8734(13) \cdot 10^{-28}\ \mathrm{m^2}$
Bohr magneton	μ_B	$= e\hbar/(2m_e) = 927{,}400\,968\,(20) \cdot 10^{-26}$ J/T
		$= 5{,}788\,381\,8066(38) \cdot 10^{-5}$ eV/T
nuclear magneton	μ_k	$= e\hbar/(2m_p) = 5{,}050\,783\,53(11) \cdot 10^{-27}$ J/T
		$= 3{,}152\,451\,2605(22) \cdot 10^{-8}$ eV/T
nuclear radius	R	$= r_0 A^{1/3};\ r_0 = (1,2 \ldots 1,4)$ fm; $1 \leq A \leq 250$: $9\ \mathrm{fm} \geq R \geq r_0$
rest energy		
atomic mass unit	$E_0(u)$	$= 931{,}494\,061(21)$ MeV
electron	$E_0(e)$	$= 0{,}510\,998\,928(11)$ MeV
proton	$E_0(p)$	$= 938{,}272\,046(21)$ MeV
neutron	$E_0(n)$	$= 939{,}565\,379(21)$ MeV
rest mass		
electron	m_e	$= 9{,}109\,382\,91(40) \cdot 10^{-31}\ \mathrm{kg} = 5{,}485\,799\,0946(22) \cdot 10^{-4}$ u
proton	m_p	$= 1{,}672\,621\,71(29) \cdot 10^{-27}\ \mathrm{kg} = 1\,836{,}152\,672\,61(85)\, m_\mathrm{e}$
		$= 1{,}007\,276\,466\,812(90)$ u
neutron	m_n	$= 1{,}674\,927\,351(74) \cdot 10^{-27}\ \mathrm{kg} = 1\,838{,}683\,659\,8(13)\, m_\mathrm{e}$
		$= 1{,}008\,664\,915\,60(55)$ u
magnetic moment		
electron	μ_e	$= -1{,}001\,159\,652\,1859(41)\ \mu_\mathrm{B}$
		$= -928{,}476\,412(80) \cdot 10^{-26}$ J/T
proton	μ_p	$= +2{,}792\,847\,356(23)\, \mu_\mathrm{k} = 1{,}410\,606\,71(12) \cdot 10^{-26}$ J/T
neutron	μ_n	$= -1{,}913\,042\,72(45)\, \mu_\mathrm{k} = 0{,}966\,236\,47(23) \cdot 10^{-26}$ J/T

21.3 Metric Prefixes

Prefix	Factor	Abbrevation	Prefix	Factor	Abbrevation
Yocto	10^{-24}	y	Deka	10^{1}	da
Zepto	10^{-21}	z	Hekto	10^{2}	h
Atto	10^{-18}	a	Kilo	10^{3}	k
Femto	10^{-15}	f	Mega	10^{6}	M
Pico	10^{-12}	p	Giga	10^{9}	G
Nano	10^{-9}	n	Tera	10^{12}	T
Mikro	10^{-6}	μ	Peta	10^{15}	P
Milli	10^{-3}	m	Exa	10^{18}	E
Zenti	10^{-2}	c	Zetta	10^{21}	Z
Dezi	10^{-1}	d	Yotta	10^{24}	Y

■ $10^3 = 1000$. ■ $10^{-3} = 0,001$. ■ 10^3 m $= 1$ km ■ $1\ \mu$m $= 10^{-6}$ m. ■ 1 nm $= 10^{-9}$ m.

Remark: The metric system is built up by adding prefixes which are the same for every kind of measure. These prefixes should be used in steps of powers with base 10 and exponent ±3: milli-, micro-, nano-; ruther than in the smaller steps hecto-, deca-, deci-. The British system, unlike the metric one, is not built up in 10's e. g.: 1 lb = 16 oz = 7000 grains.

21.4 International System of Physical Units (SI Units)

Further information about physical units see [21.8], [21.4], [21.15].

SI Base Units			
length	m	meter	
time	s	second	
mass	kg	kilogram	
thermodynamic temperature	K	kelvin	
electric current	A	ampere	
amount of substance	mol	(1 mol $=N_A$ particles, N_A = AVOGADRO–constant)	
luminous intensity	cd	candela	
Additional SI Units			
plain angle	rad	radian	$\alpha = l/r$, $1\,\text{rad} = 1\,\text{m}/1\,\text{m}$
solid angle	sr	steradian	$\Omega = S/r^2$, $1\,\text{sr} = 1\,\text{m}^2/1\,\text{m}^2$
Examples of SI derived units with special names and symbols			
frequency	Hz	Hertz	$1\,\text{Hertz} = 1/\text{s}$
force	N	Newton	$1\,\text{N} = 1\,\text{kg}\,\text{m}/\text{s}^2$
pressure, tension	Pa	Pascal	$1\,\text{Pa} = 1\,\text{N}/\text{m}^2 = 1\,\text{kg}/(\text{m}\,\text{s}^2)$
energy, work,	J	Joule	$1\,\text{J} = 1\,\text{N}\,\text{m} = 1\,\text{kg}\,\text{m}^2/\text{s}^2$
quantity of heat	kWh	kilowatt hour	$1\,\text{kWh} = 3,6\cdot 10^6\,\text{J}$
power	W	Watt	$1\,\text{W} = 1\,\text{N}\,\text{m}/\text{s} = 1\,\text{J}/\text{s} = 1\,\text{kg}\,\text{m}^2/\text{s}^3$
electric charge	C	Coulomb	$1\,\text{C} = 1\,\text{A}\,\text{s}$
electric voltage	V	Volt	$1\,\text{V} = 1\,\text{W}/\text{A} = 1\,\text{kg}\,\text{m}^2/(\text{A}\,\text{s}^3)$
electric capacitance	F	Farad	$1\,\text{F} = 1\,\text{C}/\text{V} = 1\,\text{A}^2\,\text{s}^2/\text{J} = 1\,\text{A}^2\text{s}^4/(\text{kg}\,\text{m}^2)$
electric resistance	Ω	Ohm	$1\Omega = 1\,\text{V}/\text{A} = 1\,\text{kg}\,\text{m}^2/(\text{A}^2\,\text{s}^3)$
electric conductance	S	Siemens	$1\,\text{S} = 1/\Omega = 1\,\text{A}^2\,\text{s}^3/(\text{kg}\,\text{m}^2)$
magnetic flux	Wb	Weber	$1\,\text{Wb} = 1\,\text{V}\,\text{s} = 1\,\text{kg}\,\text{m}^2/(\text{A}\,\text{s}^2)$
magnetic flux density	T	Tesla	$1\,\text{T} = 1\,\text{Wb}/\text{m}^2 = 1\,\text{kg}/(\text{A}\,\text{s}^2)$
inductance	H	Henry	$1\,\text{H} = 1\,\text{Wb}/\text{A} = 1\,\text{kg}\,\text{m}^2/(\text{A}^2\,\text{s}^2)$
luminous flux	lm	Lumen	$1\,\text{lm} = 1\,\text{cd sr}$
illuminance	lx	Lux	$1\,\text{lx} = 1\,\text{cd sr}/\text{m}^2$

Further derived SI units without special names			
speed, velocity	m/s	acceleration	m/s^2
angular velocity	rad/s	angular acceleration	rad/s^2
momentum	kg m/s	angular momentum	$\text{kg}\,\text{m}^2/\text{s}$
torque	N m	moment of inertia	$\text{kg}\,\text{m}^2$
action	J s	energy	W s
area	m^2	volume	m^3
density	kg/m^3	particle number density	m^{-3}
electric fieldstrength	V/m	magnetic fieldstrength	A/m
heat capacity	J/K	specific heat capacity	J/(K kg)
entropy	J/K	enthalpy	J

Further derived SI units with special names and symbols			
activity	Bq	Becquerel	$\mathrm{Bq} = 1\mathrm{s}^{-1}$
dose equivalent	Sv	Sievert	$\mathrm{Sv} = \mathrm{J\,kg}^{-1}$
absorbed dose	Gy	Gray	$\mathrm{Gy} = \mathrm{J\,kg}^{-1}$

Some units outside the SI accepted for use with the SI			
area	ar	Ar	$1\,\mathrm{ar} = 100\,\mathrm{m}^2$
area	b	barn	$1\,\mathrm{barn} = 10^{-28}\,\mathrm{m}^2$
volume	l	litre	$1\,\mathrm{l} = 10^{-3}\,\mathrm{m}^3$
velocity	km/h		$1\,\mathrm{km/h} = 0{,}277\,778\,\mathrm{m/s}$
mass	u	unified atomic mass unit	$1\,\mathrm{u} = 1{,}660\,5655\cdot 10^{-27}\,\mathrm{kg}$
	t	metric ton	$1\,\mathrm{t} = 1000\,\mathrm{kg}$
energy	eV	electronvolt	$1\,\mathrm{eV} = 1{,}602\,176\,565(35)$ $\cdot 10^{-19}\,\mathrm{Nm}$
focal power	dpt	diopter	$1\,\mathrm{dpt} = 1/\mathrm{m}$
pressure	bar	Bar	$1\,\mathrm{bar} = 10^5\,\mathrm{Pa}$
	mmHg	mmHg column (Torr)	$1\,\mathrm{mmHg} = 133{,}322\,\mathrm{Pa}$
plain angle	grad	$1^\circ = \pi/180\,\mathrm{rad}$	$1^\circ = 0{,}017\,453\,293\ \ldots\ \mathrm{rad}$
	minute	$1' = (1/60)^\circ = \pi/108\,00\,\mathrm{rad}$	$1' = 000\,290\,888\ \ldots\ \mathrm{rad}$
	second	$1'' = (1/60)' = \pi/648\,000\,\mathrm{rad}$	$1'' = 000\,004\,848\ \ldots\ \mathrm{rad}$
time	min	minute	$1\,\mathrm{min} = 60\,\mathrm{s}$
	h	hour	$1\,\mathrm{h} = 3{,}6\cdot 10^3\,\mathrm{s}$
	d	day	$1\,\mathrm{d} = 8{,}64\cdot 10^4\,\mathrm{s}$
	a	year	$1\,\mathrm{a} = 365\,\mathrm{d} = 8760\,\mathrm{h}$
Some units outside the SI currently accepted for use with the SI			
length	ua	astronomical unit	$1\mathrm{AE} = 149{,}597\,870\cdot 10^9\,\mathrm{m}$
	pc	parsec	$1\mathrm{pc} = 30{,}857\cdot 10^{15}\,\mathrm{m}$
	ly	light year	$1\mathrm{Lj} = 9{,}460\,447\,63\cdot 10^{15}\,\mathrm{m}$
	Å	Ångstrøm	$1\,\text{Å} = 10^{-10}\,\mathrm{m}$
	sm	nautical (intern.) mile	$1\,\mathrm{sm} = 1852\,\mathrm{m}$
volume	bbl	U.S. barrel petroleum	$1\,\mathrm{bbl} = 0{,}158\,988\,\mathrm{m}^3$
plain angle	gon	gon	$1^{\mathrm{g}} = 0{,}5\pi\cdot 10^{-2}\,\mathrm{rad}$
		gon minute	$1^{\mathrm{c}} = 0{,}5\pi\cdot 10^{-4}\,\mathrm{rad}$
		gon second	$1^{\mathrm{cc}} = 0{,}5\pi\cdot 10^{-6}\,\mathrm{rad}$
velocity	kn	knot	$1\,\mathrm{kn} = 1\mathrm{sm/h} = 0{,}5144\,\mathrm{m/s}$
energy	cal	calory	$1\,\mathrm{cal} = 4{,}1868\,\mathrm{J}$
pressure	atm	standard atmosphere	$1\,\mathrm{atm} = 1{,}013\,25\cdot 10^5\,\mathrm{Pa}$
activity	Ci	Curie	$1\,\mathrm{Ci} = 3{,}7\cdot 10^{10}\,\mathrm{Bq}$

Remark: $1\,\mathrm{Pa} = 1\mathrm{Nm}^{-2} = 10^{-5}\,\mathrm{bar} = 7{,}52\cdot 10^{-3}\,\mathrm{Torr} = 9{,}86923\cdot 10^{-6}\,\mathrm{atm}$;
$1\,\mathrm{atm} = 760\,\mathrm{Torr} = 101325\,\mathrm{Pa}$; $1\,\mathrm{Torr} = 133{,}32\,\mathrm{Pa} = 1\mathrm{mmHg}$; $1\,\mathrm{bar} = 0{,}987\,\mathrm{atm} = 760{,}06\,\mathrm{Torr}$.
About units accepted only in some EU states see [21.15].

21.5 Important Series Expansions

Function	Series Expansion	Convergence Region
	Algebraic Functions	
	Binomial Series	
$(a \pm x)^m$	After transforming to the form $a^m\left(1 \pm \frac{x}{a}\right)^m$ one gets the following series:	$\lvert x\rvert \le a$ for $m > 0$; $\lvert x\rvert < a$ for $m < 0$
	Binomial Series with Positive Exponents	
$(1 \pm x)^m$ $(m > 0)$	$1 \pm mx + \frac{m(m-1)}{2!}x^2 \pm \frac{m(m-1)(m-2)}{3!}x^3 + \cdots + (\pm 1)^n \frac{m(m-1)\ldots(m-n+1)}{n!}x^n + \cdots$	$\lvert x\rvert \le 1$
$(1 \pm x)^{\frac{1}{4}}$	$1 \pm \frac{1}{4}x - \frac{1\cdot 3}{4\cdot 8}x^2 \pm \frac{1\cdot 3\cdot 7}{4\cdot 8\cdot 12}x^3 - \frac{1\cdot 3\cdot 7\cdot 11}{4\cdot 8\cdot 12\cdot 16}x^4 \pm \cdots$	$\lvert x\rvert \le 1$
$(1 \pm x)^{\frac{1}{3}}$	$1 \pm \frac{1}{3}x - \frac{1\cdot 2}{3\cdot 6}x^2 \pm \frac{1\cdot 2\cdot 5}{3\cdot 6\cdot 9}x^3 - \frac{1\cdot 2\cdot 5\cdot 8}{3\cdot 6\cdot 9\cdot 12}x^4 \pm \cdots$	$\lvert x\rvert \le 1$
$(1 \pm x)^{\frac{1}{2}}$	$1 \pm \frac{1}{2}x - \frac{1\cdot 1}{2\cdot 4}x^2 \pm \frac{1\cdot 1\cdot 3}{2\cdot 4\cdot 6}x^3 - \frac{1\cdot 1\cdot 3\cdot 5}{2\cdot 4\cdot 6\cdot 8}x^4 \pm \cdots$	$\lvert x\rvert \le 1$
$(1 \pm x)^{\frac{3}{2}}$	$1 \pm \frac{3}{2}x + \frac{3\cdot 1}{2\cdot 4}x^2 \mp \frac{3\cdot 1\cdot 1}{2\cdot 4\cdot 6}x^3 + \frac{3\cdot 1\cdot 1\cdot 3}{2\cdot 4\cdot 6\cdot 8}x^4 \mp \cdots$	$\lvert x\rvert \le 1$
$(1 \pm x)^{\frac{5}{2}}$	$1 \pm \frac{5}{2}x + \frac{5\cdot 3}{2\cdot 4}x^2 \pm \frac{5\cdot 3\cdot 1}{2\cdot 4\cdot 6}x^3 - \frac{5\cdot 3\cdot 1\cdot 1}{2\cdot 4\cdot 6\cdot 8}x^4 \mp \cdots$	$\lvert x\rvert \le 1$
	Binomial Series with Negative Exponents	
$(1 \pm x)^{-m}$ $(m > 0)$	$1 \mp mx + \frac{m(m+1)}{2!}x^2 \mp \frac{m(m+1)(m+2)}{3!}x^3 + \cdots + (\mp 1)^n \frac{m(m+1)\ldots(m+n-1)}{n!}x^n + \cdots$	$\lvert x\rvert < 1$
$(1 \pm x)^{-\frac{1}{4}}$	$1 \mp \frac{1}{4}x + \frac{1\cdot 5}{4\cdot 8}x^2 \mp \frac{1\cdot 5\cdot 9}{4\cdot 8\cdot 12}x^3 + \frac{1\cdot 5\cdot 9\cdot 13}{4\cdot 8\cdot 12\cdot 16}x^4 \mp \cdots$	$\lvert x\rvert < 1$
$(1 \pm x)^{-\frac{1}{3}}$	$1 \mp \frac{1}{3}x + \frac{1\cdot 4}{3\cdot 6}x^2 \mp \frac{1\cdot 4\cdot 7}{3\cdot 6\cdot 9}x^3 + \frac{1\cdot 4\cdot 7\cdot 10}{3\cdot 6\cdot 9\cdot 12}x^4 \mp \cdots$	$\lvert x\rvert < 1$
$(1 \pm x)^{-\frac{1}{2}}$	$1 \mp \frac{1}{2}x + \frac{1\cdot 3}{2\cdot 4}x^2 \mp \frac{1\cdot 3\cdot 5}{2\cdot 4\cdot 6}x^3 + \frac{1\cdot 3\cdot 5\cdot 7}{2\cdot 4\cdot 6\cdot 8}x^4 \mp \cdots$	$\lvert x\rvert < 1$
$(1 \pm x)^{-1}$	$1 \mp x + x^2 \mp x^3 + x^4 \mp \cdots$	$\lvert x\rvert < 1$
$(1 \pm x)^{-\frac{3}{2}}$	$1 \mp \frac{3}{2}x + \frac{3\cdot 5}{2\cdot 4}x^2 \mp \frac{3\cdot 5\cdot 7}{2\cdot 4\cdot 6}x^3 + \frac{3\cdot 5\cdot 7\cdot 9}{2\cdot 4\cdot 6\cdot 8}x^4 \mp \cdots$	$\lvert x\rvert < 1$
$(1 \pm x)^{-2}$	$1 \mp 2x + 3x^2 \mp 4x^3 + 5x^4 \mp \cdots$	$\lvert x\rvert < 1$

Function	Series Expansion	Convergence Region
$(1 \pm x)^{-\frac{5}{2}}$	$1 \mp \frac{5}{2}x + \frac{5 \cdot 7}{2 \cdot 4}x^2 \mp \frac{5 \cdot 7 \cdot 9}{2 \cdot 4 \cdot 6}x^3 + \frac{5 \cdot 7 \cdot 9 \cdot 11}{2 \cdot 4 \cdot 6 \cdot 8}x^4 \mp \cdots$	$\|x\| < 1$
$(1 \pm x)^{-3}$	$1 \mp \frac{1}{1 \cdot 2}(2 \cdot 3x \mp 3 \cdot 4x^2 + 4 \cdot 5x^3 \mp 5 \cdot 6x^4 + \cdots)$	$\|x\| < 1$
$(1 \pm x)^{-4}$	$1 \mp \frac{1}{1 \cdot 2 \cdot 3}(2 \cdot 3 \cdot 4x \mp 3 \cdot 4 \cdot 5x^2$ $+4 \cdot 5 \cdot 6x^3 \mp 5 \cdot 6 \cdot 7x^4 + \cdots)$	$\|x\| < 1$
$(1 \pm x)^{-5}$	$1 \mp \frac{1}{1 \cdot 2 \cdot 3 \cdot 4}(2 \cdot 3 \cdot 4 \cdot 5x \mp 3 \cdot 4 \cdot 5 \cdot 6x^2$ $+4 \cdot 5 \cdot 6 \cdot 7x^3 \mp 5 \cdot 6 \cdot 7 \cdot 8x^4 + \cdots)$	$\|x\| < 1$
	Trigonometric Functions	
$\sin x$	$x - \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots + (-1)^n \frac{x^{2n+1}}{(2n+1)!} \pm \cdots$	$\|x\| < \infty$
$\sin(x+a)$	$\sin a + x \cos a - \frac{x^2 \sin a}{2!} - \frac{x^3 \cos a}{3!}$ $+\frac{x^4 \sin a}{4!} + \cdots + \frac{x^n \sin\left(a + \frac{n\pi}{2}\right)}{n!} \cdots$	$\|x\| < \infty$
$\cos x$	$1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots + (-1)^n \frac{x^{2n}}{(2n)!} \pm$	$\|x\| < \infty$
$\cos(x+a)$	$\cos a - x \sin a - \frac{x^2 \cos a}{2!} + \frac{x^3 \sin a}{3!}$ $+\frac{x^4 \cos a}{4!} - \cdots + \frac{x^n \cos\left(a + \frac{n\pi}{2}\right)}{n!} \pm \cdots$	$\|x\| < \infty$
$\tan x$	$x + \frac{1}{3}x^3 + \frac{2}{15}x^5 + \frac{17}{315}x^7 + \frac{62}{2835}x^9 + \cdots$ $+\frac{2^{2n}(2^{2n}-1)B_n}{(2n)!}x^{2n-1} + \cdots$	$\|x\| < \frac{\pi}{2}$
$\cot x$	$\frac{1}{x} - \left[\frac{x}{3} + \frac{x^3}{45} + \frac{2x^5}{945} + \frac{x^7}{4725} + \cdots\right.$ $\left.+\frac{2^{2n}B_n}{(2n)!}x^{2n-1} + \cdots\right]$	$0 < \|x\| < \pi$
$\sec x$	$1 + \frac{1}{2}x^2 + \frac{5}{24}x^4 + \frac{61}{720}x^6 + \frac{277}{8064}x^8 + \cdots$ $+\frac{E_n}{(2n)!}x^{2n} + \cdots$	$\|x\| < \frac{\pi}{2}$

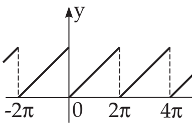
Function	Series Expansion	Convergence Region
$\operatorname{cosec} x$	$\frac{1}{x}+\frac{1}{6}x+\frac{7}{360}x^3+\frac{31}{15120}x^5+\frac{127}{604800}x^7+\cdots+\frac{2(2^{2n-1}-1)}{(2n)!}B_n x^{2n-1}$	$0<\lvert x\rvert<\pi$
	Exponential Functions	
e^x	$1+\frac{x}{1!}+\frac{x^2}{2!}+\frac{x^3}{3!}+\cdots+\frac{x^n}{n!}+\cdots$	$\lvert x\rvert<\infty$
$a^x=e^{x\ \ln a}$	$1+\frac{x\ \ln a}{1!}+\frac{(x\ \ln a)^2}{2!}+\frac{(x\ \ln a)^3}{3!}+\cdots+\frac{(x\ \ln a)^n}{n!}+\cdots$	$\lvert x\rvert<\infty$
$\frac{x}{e^x-1}$	$1-\frac{x}{2}+\frac{B_1\,x^2}{2!}-\frac{B_2\,x^4}{4!}+\frac{B_3\,x^6}{6!}-\cdots+(-1)^{n+1}\frac{B_n\,x^{2n}}{(2n)!}\pm\cdots$	$\lvert x\rvert<2\pi$
	Logarithmic Functions	
$\ln x$	$2\left[\frac{x-1}{x+1}+\frac{(x-1)^3}{3(x+1)^3}+\frac{(x-1)^5}{5(x+1)^5}+\cdots+\frac{(x-1)^{2n+1}}{(2n+1)(x+1)^{2n+1}}+\cdots\right]$	$x>0$
$\ln x$	$(x-1)-\frac{(x-1)^2}{2}+\frac{(x-1)^3}{3}-\frac{(x-1)^4}{4}+\cdots+(-1)^{n+1}\frac{(x-1)^n}{n}\pm\cdots$	$0<x\le 2$
$\ln x$	$\frac{x-1}{x}+\frac{(x-1)^2}{2x^2}+\frac{(x-1)^3}{3x^3}+\cdots+\frac{(x-1)^n}{nx^n}+\cdots$	$x>\frac{1}{2}$
$\ln(1+x)$	$x-\frac{x^2}{2}+\frac{x^3}{3}-\frac{x^4}{4}+\cdots+(-1)^{n+1}\frac{x^n}{n}\pm\cdots$	$-1<x\le 1$
$\ln(1-x)$	$-\left[x+\frac{x^2}{2}+\frac{x^3}{3}+\frac{x^4}{4}+\frac{x^5}{5}+\cdots+\frac{x^n}{n}+\cdots\right]$	$-1\le x<1$
$\ln\left(\frac{1+x}{1-x}\right)$ $=2\operatorname{Artanh} x$	$2\left[x+\frac{x^3}{3}+\frac{x^5}{5}+\frac{x^7}{7}+\cdots+\frac{x^{2n+1}}{2n+1}+\cdots\right]$	$\lvert x\rvert<1$

Function	Series Expansion	Convergence Region
$\ln\left(\frac{x+1}{x-1}\right)$ =2 Arcoth x	$2\left[\frac{1}{x}+\frac{1}{3x^3}+\frac{1}{5x^5}+\frac{1}{7x^7}+\cdots+\frac{1}{(2n+1)\,x^{2n+1}}+\cdots\right]$	$\lvert x\rvert > 1$
$\ln\lvert\sin x\rvert$	$\ln\lvert x\rvert-\frac{x^2}{6}-\frac{x^4}{180}-\frac{x^6}{2835}-\cdots-\frac{2^{2n-1}\,B_n\,x^{2n}}{n\,(2n)\,!}-\cdots$	$0<\lvert x\rvert<\pi$
$\ln\cos x$	$-\frac{x^2}{2}-\frac{x^4}{12}-\frac{x^6}{45}-\frac{17x^8}{2520}-\cdots-\frac{2^{2n-1}(2^{2n}-1)B_n\,x^{2n}}{n(2n)\,!}-\cdots$	$\lvert x\rvert<\frac{\pi}{2}$
$\ln\lvert\tan x\rvert$	$\ln\lvert x\rvert+\frac{1}{3}x^2+\frac{7}{90}x^4+\frac{62}{2835}x^6+\cdots+\frac{2^{2n}(2^{2n-1}-1)B_n}{n(2n)\,!}x^{2n}+\cdots$	$0<\lvert x\rvert<\frac{\pi}{2}$
	Inverse Trigonometric Functions	
$\arcsin x$	$x+\frac{x^3}{2\cdot 3}+\frac{1\cdot 3\,x^5}{2\cdot 4\cdot 5}+\frac{1\cdot 3\cdot 5\,x^7}{2\cdot 4\cdot 6\cdot 7}+\cdots+\frac{1\cdot 3\cdot 5\cdots(2n-1)\,x^{2n+1}}{2\cdot 4\cdot 6\cdots(2n)(2n+1)}+\cdots$	$\lvert x\rvert<1$
$\arccos x$	$\frac{\pi}{2}-\left[x+\frac{x^3}{2\cdot 3}+\frac{1\cdot 3\,x^5}{2\cdot 4\cdot 5}+\frac{1\cdot 3\cdot 5\,x^7}{2\cdot 4\cdot 6\cdot 7}+\cdots+\frac{1\cdot 3\cdot 5\cdots(2n-1)\,x^{2n+1}}{2\cdot 4\cdot 6\cdots(2n)(2n+1)}+\cdots\right]$	$\lvert x\rvert<1$
$\arctan x$	$x-\frac{x^3}{3}+\frac{x^5}{5}-\frac{x^7}{7}+\cdots+(-1)^n\frac{x^{2n+1}}{2n+1}\pm\cdots$	$\lvert x\rvert<1$
$\arctan x$	$\pm\frac{\pi}{2}-\frac{1}{x}+\frac{1}{3\,x^3}-\frac{1}{5\,x^5}+\frac{1}{7\,x^7}-\cdots+(-1)^{n+1}\frac{1}{(2n+1)\,x^{2n+1}}\pm\cdots$	$\lvert x\rvert>1$
$\operatorname{arccot} x$	$\frac{\pi}{2}-\left[x-\frac{x^3}{3}+\frac{x^5}{5}-\frac{x^7}{7}+\cdots+(-1)^n\frac{x^{2n+1}}{2n+1}\pm\cdots\right]$	$\lvert x\rvert<1$

Function	Series Expansion	Convergence Region
	Hyperbolic Functions	
$\sinh x$	$x + \frac{x^3}{3\,!} + \frac{x^5}{5\,!} + \frac{x^7}{7\,!} + \cdots + \frac{x^{2n+1}}{(2n+1)\,!} + \cdots$	$\|x\| < \infty$
$\cosh x$	$1 + \frac{x^2}{2\,!} + \frac{x^4}{4\,!} + \frac{x^6}{6\,!} + \cdots + \frac{x^{2n}}{(2n)\,!} + \cdots$	$\|x\| < \infty$
$\tanh x$	$x - \frac{1}{3}x^3 + \frac{2}{15}x^5 - \frac{17}{315}x^7 + \frac{62}{2835}x^9 - \cdots$ $+\frac{(-1)^{n+1}\,2^{2n}(2^{2n}-1)}{(2n)\,!}\,B_n x^{2n-1} \pm \cdots$	$\|x\| < \frac{\pi}{2}$
$\coth x$	$\frac{1}{x} + \frac{x}{3} - \frac{x^3}{45} + \frac{2x^5}{945} - \frac{x^7}{4725} + \cdots$ $+\frac{(-1)^{n+1}\,2^{2n}}{(2n)\,!}\,B_n x^{2n-1} \pm \cdots$	$0 < \|x\| < \pi$
$\mathrm{sech}\, x$	$1 - \frac{1}{2\,!}x^2 + \frac{5}{4\,!}x^4 - \frac{61}{6\,!}x^6 + \frac{1385}{8\,!}x^8 - \cdots$ $+\frac{(-1)^n}{(2n)\,!}\,E_n x^{2n} \pm \cdots$	$\|x\| < \frac{\pi}{2}$
$\mathrm{cosech}\, x$	$\frac{1}{x} - \frac{x}{6} + \frac{7\,x^3}{360} - \frac{31\,x^5}{15120} + \cdots$ $+\frac{2(-1)^n\,(2^{2n-1}-1)}{(2n)\,!}\,B_n x^{2n-1} + \cdots$	$0 < \|x\| < \pi$
	Area Functions	
$\mathrm{Arsinh}\, x$	$x - \frac{1}{2\cdot 3}x^3 + \frac{1\cdot 3}{2\cdot 4\cdot 5}x^5 - \frac{1\cdot 3\cdot 5}{2\cdot 4\cdot 6\cdot 7}x^7 + \cdots$ $+(-1)^n\cdot\frac{1\cdot 3\cdot 5\cdots(2n-1)}{2\cdot 4\cdot 6\cdots 2n\,(2n+1)}\,x^{2n+1} \pm \cdots$	$\|x\| < 1$
$\mathrm{Arcosh}\, x$	$\pm\left[\ln(2x) - \frac{1}{2\cdot 2x^2} - \frac{1\cdot 3}{2\cdot 4\cdot 4x^4} - \frac{1\cdot 3\cdot 5}{2\cdot 4\cdot 6x^6} - \cdots\right]$	$x > 1$
$\mathrm{Artanh}\, x$	$x + \frac{x^3}{3} + \frac{x^5}{5} + \frac{x^7}{7} + \cdots + \frac{x^{2n+1}}{2n+1} + \cdots$	$\|x\| < 1$
$\mathrm{Arcoth}\, x$	$\frac{1}{x} + \frac{1}{3x^3} + \frac{1}{5x^5} + \frac{1}{7x^7} + \cdots + \frac{1}{(2n+1)\,x^{2n+1}} + \cdots$	$\|x\| > 1$

21.6 Fourier Series

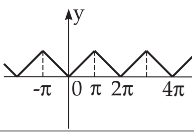
1. $y = x$ for $0 < x < 2\pi$



$$y = \pi - 2\left(\frac{\sin x}{1} + \frac{\sin 2x}{2} + \frac{\sin 3x}{3} + \cdots\right)$$

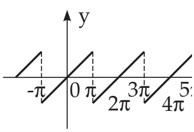
2. $y = x$ for $0 \leq x \leq \pi$

$y = 2\pi - x$ for $\pi < x \leq 2\pi$



$$y = \frac{\pi}{2} - \frac{4}{\pi}\left(\cos x + \frac{\cos 3x}{3^2} + \frac{\cos 5x}{5^2} + \cdots\right)$$

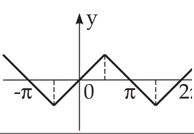
3. $y = x$ for $-\pi < x < \pi$



$$y = 2\left(\frac{\sin x}{1} - \frac{\sin 2x}{2} + \frac{\sin 3x}{3} - \cdots\right)$$

4. $y = x$ for $-\frac{\pi}{2} \leq x \leq \frac{\pi}{2}$

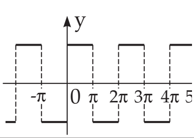
$y = \pi - x$ for $\frac{\pi}{2} \leq x \leq \frac{3\pi}{2}$



$$y = \frac{4}{\pi}\left(\sin x - \frac{\sin 3x}{3^2} + \frac{\sin 5x}{5^2} - \cdots\right)$$

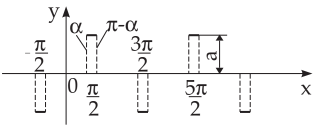
5. $y = a$ for $0 < x < \pi$

$y = -a$ for $\pi < x < 2\pi$



$$y = \frac{4a}{\pi}\left(\sin x + \frac{\sin 3x}{3} + \frac{\sin 5x}{5} + \cdots\right)$$

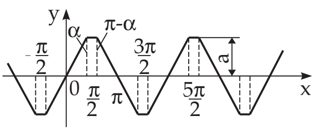
6. $y = 0$ for $0 \le x < \alpha$ and for $\pi - \alpha < x \le \pi + \alpha$ and $2\pi - \alpha < x \le 2\pi$
$y = a$ for $\alpha < x < \pi - \alpha$
$y = -a$ for $\pi + \alpha < x \le 2\pi - \alpha$



$$y = \frac{4a}{\pi}\left(\cos\alpha\sin x + \frac{1}{3}\cos 3\alpha\sin 3x + \frac{1}{5}\cos 5\alpha\sin 5x + \cdots\right)$$

---

7. $y = \frac{ax}{\alpha}$ for $-a \le x \le a$
$y = a$ for $\alpha \le x \le \pi - \alpha$
$y = \frac{a(\pi - x)}{\alpha}$ for $\pi - \alpha \le x \le \pi + \alpha$
$y = -a$ for $\pi + \alpha \le x \le 2\pi - \alpha$

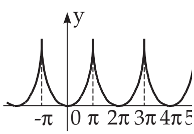


$$y = \frac{4}{\pi}\frac{a}{\alpha}\left(\sin\alpha\sin x + \frac{1}{3^2}\sin 3\alpha\sin 3x + \frac{1}{5^2}\sin 5\alpha\sin 5x + \cdots\right)$$

Especially, for $\alpha = \frac{\pi}{3}$ holds: $y = \frac{6\sqrt{3}a}{\pi^2}\left(\sin x - \frac{1}{5^2}\sin 5x + \frac{1}{7^2}\sin 7x - \frac{1}{11^2}\sin 11x + \cdots\right)$

---

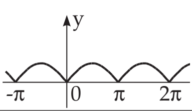
8. $y = x^2$ for $-\pi \le x \le \pi$



$$y = \frac{\pi^2}{3} - 4\left(\frac{\cos x}{1} - \frac{\cos 2x}{2^2} + \frac{\cos 3x}{3^2} - \cdots\right)$$

---

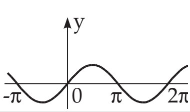
9. $y = x(\pi - x)$ for $0 \le x \le \pi$



$$y = \frac{\pi^2}{6} - \left(\frac{\cos 2x}{1^2} + \frac{\cos 4x}{2^2} + \frac{\cos 6x}{3^2} + \cdots\right)$$

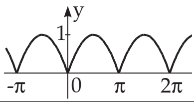
---

10. $y = x(\pi - x)$ for $0 \le x \le \pi$
$y = (\pi - x)(2\pi - x)$ for $\pi \le x \le 2\pi$



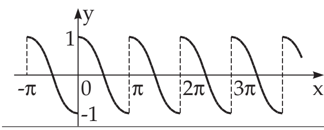
$$y = \frac{8}{\pi}\left(\sin x + \frac{1}{3^3}\sin 3x + \frac{1}{5^3}\sin 5x + \cdots\right)$$

11. $y = \sin x$ for $0 \le x \le \pi$



$$y = \frac{2}{\pi} - \frac{4}{\pi}\left(\frac{\cos 2x}{1\cdot 3} + \frac{\cos 4x}{3\cdot 5} + \frac{\cos 6x}{5\cdot 7} + \cdots\right)$$

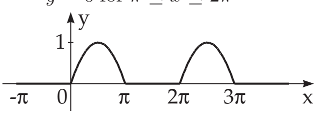
12. $y = \cos x$ for $0 < x < \pi$



$$y = \frac{4}{\pi}\left(\frac{2\sin 2x}{1\cdot 3} + \frac{4\sin 4x}{3\cdot 5} + \frac{6\sin 6x}{5\cdot 7} + \cdots\right)$$

13. $y = \sin x$ for $0 \le x \le \pi$

$y = 0$ for $\pi \le x \le 2\pi$



$$y = \frac{1}{\pi} + \frac{1}{2}\sin x - \frac{2}{\pi}\left(\frac{\cos 2x}{1\cdot 3} + \frac{\cos 4x}{3\cdot 5} + \frac{\cos 6x}{5\cdot 7} + \cdots\right)$$

14. $y = \cos ux$ for $-\pi \le x \le \pi$

$$y = \frac{2u\sin u\pi}{\pi}\left[\frac{1}{2u^2} - \frac{\cos x}{u^2-1} + \frac{\cos 2x}{u^2-4} - \frac{\cos 3x}{u^2-9} + \cdots\right]$$

(u arbitrary, but not integer number)

15. $y = \sin ux$ for $-\pi < x < \pi$

$$y = \frac{2\sin u\pi}{\pi}\left(\frac{\sin x}{1-u^2} - \frac{2\sin 2x}{4-u^2} + \frac{3\sin 3x}{9-u^2} + \cdots\right)$$

(u arbitrary, but not integer number)

16. $y = x\cos x$ for $-\pi < x < \pi$

$$y = -\frac{1}{2}\sin x + \frac{4\sin 2x}{2^2-1} - \frac{6\sin 3x}{3^2-1} + \frac{8\sin 4x}{4^2-1} - \cdots$$

17. $y = -\ln\left(2\sin\frac{x}{2}\right)$ for $0 < x \le \pi$

$$y = \cos x + \frac{1}{2}\cos 2x + \frac{1}{3}\cos 3x + \cdots$$

18. $y = \ln\left(2\cos\frac{x}{2}\right)$ for $0 \le x < \pi$

$$y = \cos x - \frac{1}{2}\cos 2x + \frac{1}{3}\cos 3x - \cdots$$

19. $y = \frac{1}{2}\ln\,\cot\frac{x}{2}$ for $0 < x < \pi$

$$y = \cos x + \frac{1}{3}\cos 3x + \frac{1}{5}\cos 5x + \cdots$$

21.7 Indefinite Integrals

(For instructions on using these tables see 8.1.1.2, **2.**, p. 482).

21.7.1 Integral Rational Functions

21.7.1.1 Integrals with $X = ax + b$

Notation: $X = ax + b$

1. $\displaystyle\int X^n\,dx = \frac{1}{a(n+1)}X^{n+1} \qquad (n \neq -1);$ (for $n = -1$ see No. 2).

2. $\displaystyle\int \frac{dx}{X} = \frac{1}{a}\ln X.$

3. $\displaystyle\int xX^n\,dx = \frac{1}{a^2(n+2)}X^{n+2} - \frac{b}{a^2(n+1)}X^{n+1}$

$(n \neq -1, \neq -2);$ (for $n = -1\,, = -2$ see No. 5 und 6).

4. $\displaystyle\int x^m X^n\,dx = \frac{1}{a^{m+1}}\int (X-b)^m X^n\,dX \qquad (n \neq -1, \neq -2, \ldots, \neq -m).$

The integral is used for $m < n$ or for integer m and fractional n; in these cases $(X-b)^m$ is expanded by the binomial theorem (see 1.1.6.4, p. 12).

5 $\displaystyle\int \frac{x\,dx}{X} = \frac{x}{a} - \frac{b}{a^2}\ln X.$

6. $\displaystyle\int \frac{x\,dx}{X^2} = \frac{b}{a^2 X} + \frac{1}{a^2}\ln X.$

7. $\displaystyle\int \frac{x\,dx}{X^3} = \frac{1}{a^2}\left(-\frac{1}{X} + \frac{b}{2X^2}\right).$

8. $\displaystyle\int \frac{x\,dx}{X^n} = \frac{1}{a^2}\left(\frac{-1}{(n-2)X^{n-2}} + \frac{b}{(n-1)X^{n-1}}\right) \qquad (n \neq 1, \neq 2).$

9. $\displaystyle\int \frac{x^2\,dx}{X} = \frac{1}{a^3}\left(\frac{1}{2}X^2 - 2bX + b^2\ln X\right).$

10. $\displaystyle\int \frac{x^2\,dx}{X^2} = \frac{1}{a^3}\left(X - 2b\ln X - \frac{b^2}{X}\right).$

11. $\displaystyle\int \frac{x^2\,dx}{X^3} = \frac{1}{a^3}\left(\ln X + \frac{2b}{X} - \frac{b^2}{2X^2}\right).$

12. $\displaystyle\int \frac{x^2\,dx}{X^n} = \frac{1}{a^3}\left[\frac{-1}{(n-3)X^{n-3}} + \frac{2b}{(n-2)X^{n-2}} - \frac{b^2}{(n-1)X^{n-1}}\right] \qquad (n \neq 1, \neq 2, \neq 3).$

13. $\displaystyle\int \frac{x^3\,dx}{X} = \frac{1}{a^4}\left(\frac{X^3}{3} - \frac{3bX^2}{2} + 3b^2X - b^3\ln X\right).$

14. $\displaystyle\int \frac{x^3\,dx}{X^2} = \frac{1}{a^4}\left(\frac{X^2}{2} - 3bX + 3b^2\ln X + \frac{b^3}{X}\right).$

15. $\displaystyle\int \frac{x^3\,dx}{X^3} = \frac{1}{a^4}\left(X - 3b\ln X - \frac{3b^2}{X} + \frac{b^3}{2X^2}\right).$

16. $\displaystyle\int \frac{x^3\,dx}{X^4} = \frac{1}{a^4}\left(\ln X + \frac{3b}{X} - \frac{3b^2}{2X^2} + \frac{b^3}{3X^3}\right).$

17. $\displaystyle\int \frac{x^3\,dx}{X^n} = \frac{1}{a^4}\left[\frac{-1}{(n-4)X^{n-4}} + \frac{3b}{(n-3)X^{n-3}} - \frac{3b^2}{(n-2)X^{n-2}} + \frac{b^3}{(n-1)X^{n-1}}\right]$

$(n \neq 1, \neq 2, \neq 3, \neq 4).$

18. $\displaystyle\int \frac{dx}{xX} = -\frac{1}{b}\ln\frac{X}{x}.$

19. $\displaystyle\int \frac{dx}{xX^2} = -\frac{1}{b^2}\left(\ln\frac{X}{x} + \frac{ax}{X}\right).$

20. $\displaystyle\int \frac{dx}{xX^3} = -\frac{1}{b^3}\left(\ln\frac{X}{x} + \frac{2ax}{X} - \frac{a^2x^2}{2X^2}\right).$

21. $\displaystyle\int \frac{dx}{xX^n} = -\frac{1}{b^n}\left[\ln\frac{X}{x} - \sum_{i=1}^{n-1}\binom{n-1}{i}\frac{(-a)^i x^i}{iX^i}\right] \qquad (n \geq 1).$

22. $\displaystyle\int \frac{dx}{x^2X} = -\frac{1}{bx} + \frac{a}{b^2}\ln\frac{X}{x}.$

23. $\displaystyle\int \frac{dx}{x^2X^2} = -a\left[\frac{1}{b^2X} + \frac{1}{ab^2x} - \frac{2}{b^3}\ln\frac{X}{x}\right].$

24. $\displaystyle\int \frac{dx}{x^2X^3} = -a\left[\frac{1}{2b^2X^2} + \frac{2}{b^3X} + \frac{1}{ab^3x} - \frac{3}{b^4}\ln\frac{X}{x}\right].$

25. $\displaystyle\int \frac{dx}{x^2X^n} = -\frac{1}{b^{n+1}}\left[-\sum_{i=2}^{n}\binom{n}{i}\frac{(-a)^i x^{i-1}}{(i-1)X^{i-1}} + \frac{X}{x} - na\ln\frac{X}{x}\right] \qquad (n \geq 2).$

26. $\displaystyle\int \frac{dx}{x^3X} = -\frac{1}{b^3}\left[a^2\ln\frac{X}{x} - \frac{2aX}{x} + \frac{X^2}{2x^2}\right].$

27. $\displaystyle\int \frac{dx}{x^3X^2} = -\frac{1}{b^4}\left[3a^2\ln\frac{X}{x} + \frac{a^3x}{X} + \frac{X^2}{2x^2} - \frac{3aX}{x}\right].$

28. $\displaystyle\int \frac{dx}{x^3X^3} = -\frac{1}{b^5}\left[6a^2\ln\frac{X}{x} + \frac{4a^3x}{X} - \frac{a^4x^2}{2X^2} + \frac{X^2}{2x^2} - \frac{4aX}{x}\right].$

29. $\displaystyle\int \frac{dx}{x^3X^n} = -\frac{1}{b^{n+2}}\left[-\sum_{i=3}^{n+1}\binom{n+1}{i}\frac{(-a)^i x^{i-2}}{(i-2)X^{i-2}} + \frac{a^2X^2}{2x^2} - \frac{(n+1)aX}{x}\right.$

$\displaystyle\left. + \frac{n(n+1)a^2}{2}\ln\frac{X}{x}\right] \qquad (n \geq 3).$

30. $\displaystyle\int \frac{dx}{x^mX^n} = -\frac{1}{b^{m+n-1}}\sum_{i=0}^{m+n-2}\binom{m+n-2}{i}\frac{X^{m-i-1}(-a)^i}{(m-i-1)x^{m-i-1}}.$

If the denominators of the terms behind the sum vanish, then such terms should be replaced by

$$\binom{m+n-2}{m-1}(-a)^{m-1}\ln\frac{X}{x}.$$

Notation: $\Delta = bf - ag$

31. $\displaystyle\int \frac{ax+b}{fx+g}\,dx = \frac{ax}{f} + \frac{\Delta}{f^2}\ln(fx+g).$

32. $\displaystyle\int \frac{dx}{(ax+b)(fx+g)} = \frac{1}{\Delta}\ln\frac{fx+g}{ax+b} \qquad (\Delta \neq 0).$

33. $\displaystyle\int \frac{x\,dx}{(ax+b)(fx+g)} = \frac{1}{\Delta}\left[\frac{b}{a}\ln(ax+b) - \frac{g}{f}\ln(fx+g)\right] \qquad (\Delta \neq 0).$

34. $\displaystyle\int \frac{dx}{(ax+b)^2(fx+g)} = \frac{1}{\Delta}\left(\frac{1}{ax+b} + \frac{f}{\Delta}\ln\frac{fx+g}{ax+b}\right) \qquad (\Delta \neq 0).$

35. $\displaystyle\int \frac{x\,dx}{(a+x)(b+x)^2} = \frac{b}{(a-b)(b+x)} - \frac{a}{(a-b)^2}\ln\frac{a+x}{b+x} \qquad (a \neq b).$

36. $\displaystyle\int \frac{x^2\,dx}{(a+x)(b+x)^2} = \frac{b^2}{(b-a)(b+x)} + \frac{a^2}{(b-a)^2}\ln(a+x) + \frac{b^2-2ab}{(b-a)^2}\ln(b+x) \qquad (a \neq b).$

37. $\displaystyle\int \frac{dx}{(a+x)^2(b+x)^2} = \frac{-1}{(a-b)^2}\left(\frac{1}{a+x} + \frac{1}{b+x}\right) + \frac{2}{(a-b)^3}\ln\frac{a+x}{b+x} \qquad (a \neq b).$

38. $\displaystyle\int \frac{x\,dx}{(a+x)^2(b+x)^2} = \frac{1}{(a-b)^2}\left(\frac{a}{a+x} + \frac{b}{b+x}\right) - \frac{a+b}{(a-b)^3}\ln\frac{a+x}{b+x} \qquad (a \neq b).$

39. $\displaystyle\int \frac{x^2\,dx}{(a+x)^2(b+x)^2} = \frac{-1}{(a-b)^2}\left(\frac{a^2}{a+x} + \frac{b^2}{b+x}\right) + \frac{2ab}{(a-b)^3}\ln\frac{a+x}{b+x} \qquad (a \neq b).$

21.7.1.2 Integrals with $X = ax^2 + bx + c$

Notation: $X = ax^2 + bx + c\,;\ \Delta = 4ac - b^2$

40. $$\begin{aligned}\int \frac{dx}{X} &= \frac{2}{\sqrt{\Delta}}\arctan\frac{2ax+b}{\sqrt{\Delta}} && (\text{for } \Delta > 0),\\ &= -\frac{2}{\sqrt{-\Delta}}\operatorname{Artanh}\frac{2ax+b}{\sqrt{-\Delta}} && (\text{for } \Delta < 0),\\ &= \frac{1}{\sqrt{-\Delta}}\ln\frac{2ax+b-\sqrt{-\Delta}}{2ax+b+\sqrt{-\Delta}} && (\text{for } \Delta < 0).\end{aligned}$$

41. $\displaystyle\int \frac{dx}{X^2} = \frac{2ax+b}{\Delta X} + \frac{2a}{\Delta}\int\frac{dx}{X}$ (see No. 40).

42. $\displaystyle\int \frac{dx}{X^3} = \frac{2ax+b}{\Delta}\left(\frac{1}{2X^2} + \frac{3a}{\Delta X}\right) + \frac{6a^2}{\Delta^2}\int\frac{dx}{X}$ (see No. 40).

43. $\displaystyle\int \frac{dx}{X^n} = \frac{2ax+b}{(n-1)\Delta X^{n-1}} + \frac{(2n-3)2a}{(n-1)\Delta}\int\frac{dx}{X^{n-1}}.$

44. $\displaystyle\int \frac{x\,dx}{X} = \frac{1}{2a}\ln X - \frac{b}{2a}\int \frac{dx}{X}$ (see No. 40).

45. $\displaystyle\int \frac{x\,dx}{X^2} = -\frac{bx+2c}{\Delta X} - \frac{b}{\Delta}\int \frac{dx}{X}$ (see No. 40).

46. $$\int \frac{x\,dx}{X^n} = -\frac{bx+2c}{(n-1)\Delta X^{n-1}} - \frac{b(2n-3)}{(n-1)\Delta}\int \frac{dx}{X^{n-1}}.$$

47. $\displaystyle\int \frac{x^2\,dx}{X} = \frac{x}{a} - \frac{b}{2a^2}\ln X + \frac{b^2-2ac}{2a^2}\int \frac{dx}{X}$ (see No. 40).

48. $\displaystyle\int \frac{x^2\,dx}{X^2} = \frac{(b^2-2ac)x+bc}{a\Delta X} + \frac{2c}{\Delta}\int \frac{dx}{X}$ (see No. 40).

49. $\displaystyle\int \frac{x^2\,dx}{X^n} = \frac{-x}{(2n-3)aX^{n-1}} + \frac{c}{(2n-3)a}\int \frac{dx}{X^n} - \frac{(n-2)b}{(2n-3)a}\int \frac{x\,dx}{X^n}$ (see No. 43 and 46).

50. $$\int \frac{x^m\,dx}{X^n} = -\frac{x^{m-1}}{(2n-m-1)aX^{n-1}} + \frac{(m-1)c}{(2n-m-1)a}\int \frac{x^{m-2}\,dx}{X^n} - \frac{(n-m)b}{(2n-m-1)a}\int \frac{x^{m-1}\,dx}{X^n}$$ $(m \neq 2n-1)$; (for $m = 2n-1$ see No. 51).

51. $$\int \frac{x^{2n-1}\,dx}{X^n} = \frac{1}{a}\int \frac{x^{2n-3}\,dx}{X^{n-1}} - \frac{c}{a}\int \frac{x^{2n-3}\,dx}{X^n} - \frac{b}{a}\int \frac{x^{2n-2}\,dx}{X^n}.$$

52. $\displaystyle\int \frac{dx}{xX} = \frac{1}{2c}\ln\frac{x^2}{X} - \frac{b}{2c}\int \frac{dx}{X}$ (see No. 40).

53. $$\int \frac{dx}{xX^n} = \frac{1}{2c(n-1)X^{n-1}} - \frac{b}{2c}\int \frac{dx}{X^n} + \frac{1}{c}\int \frac{dx}{xX^{n-1}}.$$

54. $\displaystyle\int \frac{dx}{x^2X} = \frac{b}{2c^2}\ln\frac{X}{x^2} - \frac{1}{cx} + \left(\frac{b^2}{2c^2} - \frac{a}{c}\right)\int \frac{dx}{X}$ (see No. 40).

55. $$\int \frac{dx}{x^mX^n} = -\frac{1}{(m-1)cx^{m-1}X^{n-1}} - \frac{(2n+m-3)a}{(m-1)c}\int \frac{dx}{x^{m-2}X^n} - \frac{(n+m-2)b}{(m-1)c}\int \frac{dx}{x^{m-1}X^n}$$ $(m > 1)$.

56. $$\int \frac{dx}{(fx+g)X} = \frac{1}{2(cf^2-gbf+g^2a)}\left[f\ln\frac{(fx+g)^2}{X}\right] + \frac{2ga-bf}{2(cf^2-gbf+g^2a)}\int \frac{dx}{X}$$ (see No. 40).

21.7.1.3 Integrals with $X = a^2 \pm x^2$

Notation: $X = a^2 \pm x^2$,

$$Y = \begin{cases} \arctan\dfrac{x}{a} & \text{for the “+” sign,} \\ \operatorname{Artanh}\dfrac{x}{a} = \dfrac{1}{2}\ln\dfrac{a+x}{a-x} & \text{for the “−” sign and } |x| < a, \\ \operatorname{Arcoth}\dfrac{x}{a} = \dfrac{1}{2}\ln\dfrac{x+a}{x-a} & \text{for the “−” sign and } |x| > a. \end{cases}$$

If there is a double sign in a formula, then the upper one belongs to $X = a^2 + x^2$, and the lower one to $X = a^2 - x^2$, $a > 0$.

57. $\displaystyle\int \frac{dx}{X} = \frac{1}{a}Y.$

58. $\displaystyle\int \frac{dx}{X^2} = \frac{x}{2a^2X} + \frac{1}{2a^3}Y.$

59. $\displaystyle\int \frac{dx}{X^3} = \frac{x}{4a^2X^2} + \frac{3x}{8a^4X} + \frac{3}{8a^5}Y.$

60. $\displaystyle\int \frac{dx}{X^{n+1}} = \frac{x}{2na^2X^n} + \frac{2n-1}{2na^2}\int \frac{dx}{X^n}.$

61. $\displaystyle\int \frac{x\,dx}{X} = \pm\frac{1}{2}\ln X.$

62. $\displaystyle\int \frac{x\,dx}{X^2} = \mp\frac{1}{2X}.$

63. $\displaystyle\int \frac{x\,dx}{X^3} = \mp\frac{1}{4X^2}.$

64. $\displaystyle\int \frac{x\,dx}{X^{n+1}} = \mp\frac{1}{2nX^n} \qquad (n \neq 0).$

65. $\displaystyle\int \frac{x^2\,dx}{X} = \pm x \mp aY.$

66. $\displaystyle\int \frac{x^2\,dx}{X^2} = \mp\frac{x}{2X} \pm \frac{1}{2a}Y.$

67. $\displaystyle\int \frac{x^2\,dx}{X^3} = \mp\frac{x}{4X^2} \pm \frac{x}{8a^2X} \pm \frac{1}{8a^3}Y.$

68. $\displaystyle\int \frac{x^2\,dx}{X^{n+1}} = \mp\frac{x}{2nX^n} \pm \frac{1}{2n}\int \frac{dx}{X^n} \qquad (n \neq 0).$

69. $\displaystyle\int \frac{x^3\,dx}{X} = \pm\frac{x^2}{2} - \frac{a^2}{2}\ln X.$

70. $\displaystyle\int \frac{x^3\,dx}{X^2} = \frac{a^2}{2X} + \frac{1}{2}\ln X.$

71. $\displaystyle\int \frac{x^3\,dx}{X^3} = -\frac{1}{2X} + \frac{a^2}{4X^2}.$

72. $\displaystyle\int \frac{x^3\,dx}{X^{n+1}} = -\frac{1}{2(n-1)X^{n-1}} + \frac{a^2}{2nX^n} \qquad (n > 1).$

73. $\displaystyle\int \frac{dx}{xX} = \frac{1}{2a^2}\ln\frac{x^2}{X}.$

74. $\displaystyle\int \frac{dx}{xX^2} = \frac{1}{2a^2X} + \frac{1}{2a^4}\ln\frac{x^2}{X}.$

75. $\displaystyle\int \frac{dx}{xX^3} = \frac{1}{4a^2X^2} + \frac{1}{2a^4X} + \frac{1}{2a^6}\ln\frac{x^2}{X}.$

76. $\displaystyle\int \frac{dx}{x^2X} = -\frac{1}{a^2x} \mp \frac{1}{a^3}Y.$

77. $\displaystyle\int \frac{dx}{x^2X^2} = -\frac{1}{a^4x} \mp \frac{x}{2a^4X} \mp \frac{3}{2a^5}Y.$

78. $\displaystyle\int \frac{dx}{x^2X^3} = -\frac{1}{a^6x} \mp \frac{x}{4a^4X^2} \mp \frac{7x}{8a^6X} \mp \frac{15}{8a^7}Y.$

79. $\displaystyle\int \frac{dx}{x^3X} = -\frac{1}{2a^2x^2} \mp \frac{1}{2a^4}\ln\frac{x^2}{X}.$

80. $\displaystyle\int \frac{dx}{x^3X^2} = -\frac{1}{2a^4x^2} \mp \frac{1}{2a^4X} \mp \frac{1}{a^6}\ln\frac{x^2}{X}.$

81. $\displaystyle\int \frac{dx}{x^3X^3} = -\frac{1}{2a^6x^2} \mp \frac{1}{a^6X} \mp \frac{1}{4a^4X^2} \mp \frac{3}{2a^8}\ln\frac{x^2}{X}.$

82. $\displaystyle\int \frac{dx}{(b+cx)X} = \frac{1}{a^2c^2 \pm b^2}\left[c\ln(b+cx) - \frac{c}{2}\ln X \pm \frac{b}{a}Y\right].$

21.7.1.4 Integrals with $X = a^3 \pm x^3$

Notation: $a^3 \pm x^3 = X$; if there is a double sign in a formula, the upper sign belongs to $X = a^3 + x^3$, the lower one to $X = a^3 - x^3$.

83. $\displaystyle\int \frac{dx}{X} = \pm\frac{1}{6a^2}\ln\frac{(a\pm x)^2}{a^2 \mp ax + x^2} + \frac{1}{a^2\sqrt{3}}\arctan\frac{2x \mp a}{a\sqrt{3}}.$

84. $\displaystyle\int \frac{dx}{X^2} = \frac{x}{3a^3X} + \frac{2}{3a^3}\int\frac{dx}{X}$ (see No. 83).

85. $\displaystyle\int \frac{x\,dx}{X} = \frac{1}{6a}\ln\frac{a^2 \mp ax + x^2}{(a\pm x)^2} \pm \frac{1}{a\sqrt{3}}\arctan\frac{2x \mp a}{a\sqrt{3}}.$

86. $\displaystyle\int \frac{x\,dx}{X^2} = \frac{x^2}{3a^3X} + \frac{1}{3a^3}\int\frac{x\,dx}{X}$ (see No. 85).

87. $\displaystyle\int \frac{x^2\,dx}{X} = \pm\frac{1}{3}\ln X.$

88. $\displaystyle\int \frac{x^2\,dx}{X^2} = \mp\frac{1}{3X}.$

89. $\displaystyle\int \frac{x^3\,dx}{X} = \pm x \mp a^3\int\frac{dx}{X}$ (see No. 83).

90. $\displaystyle\int \frac{x^3\,dx}{X^2} = \mp\frac{x}{3X} \pm \frac{1}{3}\int\frac{dx}{X}$ (see No. 83).

91. $\displaystyle\int \frac{dx}{xX} = \frac{1}{3a^3}\ln\frac{x^3}{X}.$

92. $\displaystyle\int \frac{dx}{xX^2} = \frac{1}{3a^3X} + \frac{1}{3a^6}\ln\frac{x^3}{X}.$

93. $\displaystyle\int \frac{dx}{x^2X} = -\frac{1}{a^3x} \mp \frac{1}{a^3}\int\frac{x\,dx}{X}$ (see No. 85).

94. $\displaystyle\int \frac{dx}{x^2X^2} = -\frac{1}{a^6x} \mp \frac{x^2}{3a^6X} \mp \frac{4}{3a^6}\int\frac{x\,dx}{X}$ (see No. 85).

95. $\int \frac{dx}{x^3 X} = -\frac{1}{2a^3 x^2} \mp \frac{1}{a^3} \int \frac{dx}{X}$ (see No. 83).

96. $\int \frac{dx}{x^3 X^2} = -\frac{1}{2a^6 x^2} \mp \frac{x}{3a^6 X} \mp \frac{5}{3a^6} \int \frac{dx}{X}$ (see No. 83).

21.7.1.5 Integrals with $X = a^4 + x^4$

97. $\int \frac{dx}{a^4 + x^4} = \frac{1}{4a^3\sqrt{2}} \ln \frac{x^2 + ax\sqrt{2} + a^2}{x^2 - ax\sqrt{2} + a^2} + \frac{1}{2a^3\sqrt{2}} \arctan \frac{ax\sqrt{2}}{a^2 - x^2}.$

98. $\int \frac{x\,dx}{a^4 + x^4} = \frac{1}{2a^2} \arctan \frac{x^2}{a^2}.$

99. $\int \frac{x^2\,dx}{a^4 + x^4} = -\frac{1}{4a\sqrt{2}} \ln \frac{x^2 + ax\sqrt{2} + a^2}{x^2 - ax\sqrt{2} + a^2} + \frac{1}{2a\sqrt{2}} \arctan \frac{ax\sqrt{2}}{a^2 - x^2}.$

100. $\int \frac{x^3\,dx}{a^4 + x^4} = \frac{1}{4} \ln(a^4 + x^4).$

21.7.1.6 Integrals with $X = a^4 - x^4$

101. $\int \frac{dx}{a^4 - x^4} = \frac{1}{4a^3} \ln \frac{a + x}{a - x} + \frac{1}{2a^3} \arctan \frac{x}{a}.$

102. $\int \frac{x\,dx}{a^4 - x^4} = \frac{1}{4a^3} \ln \frac{a^2 + x^2}{a^2 - x^2}.$

103. $\int \frac{x^2\,dx}{a^4 - x^4} = \frac{1}{4a} \ln \frac{a + x}{a - x} - \frac{1}{2a} \arctan \frac{x}{a}.$

104. $\int \frac{x^3\,dx}{a^4 - x^4} = -\frac{1}{4} \ln(a^4 - x^4).$

21.7.1.7 Some Cases of Partial Fraction Decomposition

105. $\frac{1}{(a + bx)(f + gx)} \equiv \frac{1}{fb - ag} \left(\frac{b}{a + bx} - \frac{g}{f + gx} \right).$

106. $\frac{1}{(x + a)(x + b)(x + c)} \equiv \frac{A}{x + a} + \frac{B}{x + b} + \frac{C}{x + c},$ where it holds

$A = \frac{1}{(b - a)(c - a)},\ B = \frac{1}{(a - b)(c - b)},\ C = \frac{1}{(a - c)(b - c)}.$

107. $\frac{1}{(x + a)(x + b)(x + c)(x + d)} \equiv \frac{A}{x + a} + \frac{B}{x + b} + \frac{C}{x + c} + \frac{D}{x + d},$ where it holds

$A = \frac{1}{(b - a)(c - a)(d - a)},\ B = \frac{1}{(a - b)(c - b)(d - b)}$ etc.

108. $\frac{1}{(a + bx^2)(f + gx^2)} \equiv \frac{1}{fb - ag} \left(\frac{b}{a + bx^2} - \frac{g}{f + gx^2} \right).$

21.7.2 Integrals of Irrational Functions

21.7.2.1 Integrals with $\sqrt{x}$ and $a^2 \pm b^2 x$

Notation:

$$X = a^2 \pm b^2 x,\; Y = \begin{cases} \arctan \dfrac{b\sqrt{x}}{a} & \text{for the sign "+"}, \\ \dfrac{1}{2}\ln \dfrac{a + b\sqrt{x}}{a - b\sqrt{x}} & \text{for the sign "−"}. \end{cases}$$

If there is a double sign in a formula, then the upper one belongs to $X = a^2 + b^2x$, the lower one to $X = a^2 - b^2x$.

109. $\displaystyle\int \frac{\sqrt{x}\,dx}{X} = \pm\frac{2\sqrt{x}}{b^2} \mp \frac{2a}{b^3}Y.$

110. $\displaystyle\int \frac{\sqrt{x^3}\,dx}{X} = \pm\frac{2}{3}\frac{\sqrt{x^3}}{b^2} - \frac{2a^2\sqrt{x}}{b^4} + \frac{2a^3}{b^5}Y.$

111. $\displaystyle\int \frac{\sqrt{x}\,dx}{X^2} = \mp\frac{\sqrt{x}}{b^2X} \pm \frac{1}{ab^3}Y.$

112. $\displaystyle\int \frac{\sqrt{x^3}\,dx}{X^2} = \pm\frac{2\sqrt{x^3}}{b^2X} + \frac{3a^2\sqrt{x}}{b^4X} - \frac{3a}{b^5}Y.$

113. $\displaystyle\int \frac{dx}{X\sqrt{x}} = \frac{2}{ab}Y.$

114. $\displaystyle\int \frac{dx}{X\sqrt{x^3}} = -\frac{2}{a^2\sqrt{x}} \mp \frac{2b}{a^3}Y.$

115. $\displaystyle\int \frac{dx}{X^2\sqrt{x}} = \frac{\sqrt{x}}{a^2X} + \frac{1}{a^3b}Y.$

116. $\displaystyle\int \frac{dx}{X^2\sqrt{x^3}} = -\frac{2}{a^2X\sqrt{x}} \mp \frac{3b^2\sqrt{x}}{a^4X} \mp \frac{3b}{a^5}Y.$

21.7.2.2 Other Integrals with $\sqrt{x}$

117. $\displaystyle\int \frac{\sqrt{x}\,dx}{a^4 + x^2} = -\frac{1}{2a\sqrt{2}}\ln\frac{x + a\sqrt{2x} + a^2}{x - a\sqrt{2x} + a^2} + \frac{1}{a\sqrt{2}}\arctan\frac{a\sqrt{2x}}{a^2 - x}.$

118. $\displaystyle\int \frac{dx}{(a^4 + x^2)\sqrt{x}} = \frac{1}{2a^3\sqrt{2}}\ln\frac{x + a\sqrt{2x} + a^2}{x - a\sqrt{2x} + a^2} + \frac{1}{a^3\sqrt{2}}\arctan\frac{a\sqrt{2x}}{a^2 - x}.$

119. $\displaystyle\int \frac{\sqrt{x}\,dx}{a^4 - x^2} = \frac{1}{2a}\ln\frac{a + \sqrt{x}}{a - \sqrt{x}} - \frac{1}{a}\arctan\frac{\sqrt{x}}{a}.$

120. $\displaystyle\int \frac{dx}{(a^4 - x^2)\sqrt{x}} = \frac{1}{2a^3}\ln\frac{a + \sqrt{x}}{a - \sqrt{x}} + \frac{1}{a^3}\arctan\frac{\sqrt{x}}{a}.$

21.7.2.3 Integrals with $\sqrt{ax+b}$

Notation: $X = ax + b$

121. $\int \sqrt{X}\,dx = \frac{2}{3a}\sqrt{X^3}.$

122. $\int x\sqrt{X}\,dx = \frac{2(3ax-2b)\sqrt{X^3}}{15a^2}.$

123. $\int x^2\sqrt{X}\,dx = \frac{2(15a^2x^2-12abx+8b^2)\sqrt{X^3}}{105a^3}.$

124. $\int \frac{dx}{\sqrt{X}} = \frac{2\sqrt{X}}{a}.$

125. $\int \frac{x\,dx}{\sqrt{X}} = \frac{2(ax-2b)}{3a^2}\sqrt{X}.$

126. $\int \frac{x^2\,dx}{\sqrt{X}} = \frac{2(3a^2x^2-4abx+8b^2)\sqrt{X}}{15a^3}.$

127. $\int \frac{dx}{x\sqrt{X}} = \begin{cases} -\frac{2}{\sqrt{b}}\operatorname{Arcoth}\sqrt{\frac{X}{b}} = \frac{1}{\sqrt{b}}\ln\frac{\sqrt{X}-\sqrt{b}}{\sqrt{X}+\sqrt{b}} & \text{for } b>0, \\ \frac{2}{\sqrt{-b}}\arctan\sqrt{\frac{X}{-b}} & \text{for } b<0. \end{cases}$

128. $\int \frac{\sqrt{X}}{x}dx = 2\sqrt{X} + b\int\frac{dx}{x\sqrt{X}}$ (see No. 127).

129. $\int \frac{dx}{x^2\sqrt{X}} = -\frac{\sqrt{X}}{bx} - \frac{a}{2b}\int\frac{dx}{x\sqrt{X}}$ (see No. 127).

130. $\int \frac{\sqrt{X}}{x^2}dx = -\frac{\sqrt{X}}{x} + \frac{a}{2}\int\frac{dx}{x\sqrt{X}}$ (see No. 127).

131. $\int \frac{dx}{x^n\sqrt{X}} = -\frac{\sqrt{X}}{(n-1)bx^{n-1}} - \frac{(2n-3)a}{(2n-2)b}\int\frac{dx}{x^{n-1}\sqrt{X}}.$

132. $\int \sqrt{X^3}\,dx = \frac{2\sqrt{X^5}}{5a}.$

133. $\int x\sqrt{X^3}\,dx = \frac{2}{35a^2}\left(5\sqrt{X^7} - 7b\sqrt{X^5}\right).$

134. $\int x^2\sqrt{X^3}\,dx = \frac{2}{a^3}\left(\frac{\sqrt{X^9}}{9} - \frac{2b\sqrt{X^7}}{7} + \frac{b^2\sqrt{X^5}}{5}\right).$

135. $\int \frac{\sqrt{X^3}}{x}\,dx = \frac{2\sqrt{X^3}}{3} + 2b\sqrt{X} + b^2\int\frac{dx}{x\sqrt{X}}$ (see No. 127).

136. $\int \frac{x\,dx}{\sqrt{X^3}} = \frac{2}{a^2}\left(\sqrt{X} + \frac{b}{\sqrt{X}}\right).$

137. $\displaystyle\int \frac{x^2\,dx}{\sqrt{X^3}} = \frac{2}{a^3}\left(\frac{\sqrt{X^3}}{3} - 2b\sqrt{X} - \frac{b^2}{\sqrt{X}}\right).$

138. $\displaystyle\int \frac{dx}{x\sqrt{X^3}} = \frac{2}{b\sqrt{X}} + \frac{1}{b}\int \frac{dx}{x\sqrt{X}}$ (see No. 127).

139. $\displaystyle\int \frac{dx}{x^2\sqrt{X^3}} = -\frac{1}{bx\sqrt{X}} - \frac{3a}{b^2\sqrt{X}} - \frac{3a}{2b^2}\int \frac{dx}{x\sqrt{X}}$ (see No. 127).

140. $\displaystyle\int X^{\pm n/2}\,dx = \frac{2X^{(2\pm n)/2}}{a(2\pm n)}.$

141. $\displaystyle\int xX^{\pm n/2}\,dx = \frac{2}{a^2}\left(\frac{X^{(4\pm n)/2}}{4\pm n} - \frac{bX^{(2\pm n)/2}}{2\pm n}\right).$

142. $\displaystyle\int x^2X^{\pm n/2}\,dx = \frac{2}{a^3}\left(\frac{X^{(6\pm n)/2}}{6\pm n} - \frac{2bX^{(4\pm n)/2}}{4\pm n} + \frac{b^2X^{(2\pm n)/2}}{2\pm n}\right).$

143. $\displaystyle\int \frac{X^{n/2}\,dx}{x} = \frac{2X^{n/2}}{n} + b\int \frac{X^{(n-2)/2}}{x}dx.$

144. $\displaystyle\int \frac{dx}{xX^{n/2}} = \frac{2}{(n-2)bX^{(n-2)/2}} + \frac{1}{b}\int \frac{dx}{xX^{(n-2)/2}}.$

145. $\displaystyle\int \frac{dx}{x^2X^{n/2}} = -\frac{1}{bxX^{(n-2)/2}} - \frac{na}{2b}\int \frac{dx}{xX^{n/2}}.$

21.7.2.4 Integrals with $\sqrt{ax+b}$ and $\sqrt{fx+g}$

Notation: $X = ax+b$, $Y = fx+g$, $\Delta = bf - ag$

146. $$\int \frac{dx}{\sqrt{XY}} = \begin{cases} -\dfrac{2}{\sqrt{-af}}\arctan\sqrt{-\dfrac{fX}{aY}} & \text{for } af < 0\,, \\ \dfrac{2}{\sqrt{af}}\operatorname{Artanh}\sqrt{\dfrac{fX}{aY}} & \text{for } af > 0\,, \\ \dfrac{2}{\sqrt{af}}\ln\left(\sqrt{aY} + \sqrt{fX}\right) & \text{for } af > 0\,. \end{cases}$$

147. $\displaystyle\int \frac{x\,dx}{\sqrt{XY}} = \frac{\sqrt{XY}}{af} - \frac{ag+bf}{2af}\int \frac{dx}{\sqrt{XY}}$ (see No. 146).

148. $\displaystyle\int \frac{dx}{\sqrt{X}\sqrt{Y^3}} = -\frac{2\sqrt{X}}{\Delta\sqrt{Y}}.$

149. $$\int \frac{dx}{Y\sqrt{X}} = \begin{cases} \dfrac{2}{\sqrt{-\Delta f}}\arctan\dfrac{f\sqrt{X}}{\sqrt{-\Delta f}} & \text{for } \Delta f < 0\,, \\ \dfrac{1}{\sqrt{\Delta f}}\ln\dfrac{f\sqrt{X}-\sqrt{\Delta f}}{f\sqrt{X}+\sqrt{\Delta f}} & \text{for } \Delta f > 0\,. \end{cases}$$

150. $\displaystyle\int \sqrt{XY}\,dx = \frac{\Delta + 2aY}{4af}\sqrt{XY} - \frac{\Delta^2}{8af}\int \frac{dx}{\sqrt{XY}}$ (see No. 146).

151. $\int \sqrt{\frac{Y}{X}} dx = \frac{1}{a}\sqrt{XY} - \frac{\Delta}{2a}\int \frac{dx}{\sqrt{XY}}$ (see No. 146).

152. $\int \frac{\sqrt{X}\, dx}{Y} = \frac{2\sqrt{X}}{f} + \frac{\Delta}{f}\int \frac{dx}{Y\sqrt{X}}$ (see No. 149).

153. $\int \frac{Y^n\, dx}{\sqrt{X}} = \frac{2}{(2n+1)a}\left(\sqrt{X}Y^n - n\Delta \int \frac{Y^{n-1}\, dx}{\sqrt{X}}\right).$

154. $\int \frac{dx}{\sqrt{X}Y^n} = -\frac{1}{(n-1)\Delta}\left\{\frac{\sqrt{X}}{Y^{n-1}} + \left(n - \frac{3}{2}\right) a \int \frac{dx}{\sqrt{X}Y^{n-1}}\right\}.$

155. $\int \sqrt{X}Y^n\, dx = \frac{1}{(2n+3)f}\left(2\sqrt{X}Y^{n+1} + \Delta \int \frac{Y^n\, dx}{\sqrt{X}}\right)$ (see No. 153).

156. $\int \frac{\sqrt{X}\, dx}{Y^n} = \frac{1}{(n-1)f}\left(-\frac{\sqrt{X}}{Y^{n-1}} + \frac{a}{2}\int \frac{dx}{\sqrt{X}Y^{n-1}}\right).$

21.7.2.5 Integrals with $\sqrt{a^2 - x^2}$

Notation: $X = a^2 - x^2$

157. $\int \sqrt{X}\, dx = \frac{1}{2}\left(x\sqrt{X} + a^2 \arcsin \frac{x}{a}\right).$

158. $\int x\sqrt{X}\, dx = -\frac{1}{3}\sqrt{X^3}.$

159. $\int x^2\sqrt{X}\, dx = -\frac{x}{4}\sqrt{X^3} + \frac{a^2}{8}\left(x\sqrt{X} + a^2 \arcsin \frac{x}{a}\right).$

160. $\int x^3\sqrt{X}\, dx = \frac{\sqrt{X^5}}{5} - a^2\frac{\sqrt{X^3}}{3}.$

161. $\int \frac{\sqrt{X}}{x} dx = \sqrt{X} - a \ln \frac{a + \sqrt{X}}{x}.$

162. $\int \frac{\sqrt{X}}{x^2} dx = -\frac{\sqrt{X}}{x} - \arcsin \frac{x}{a}.$

163. $\int \frac{\sqrt{X}}{x^3} dx = -\frac{\sqrt{X}}{2x^2} + \frac{1}{2a} \ln \frac{a + \sqrt{X}}{x}.$

164. $\int \frac{dx}{\sqrt{X}} = \arcsin \frac{x}{a}.$

165. $\int \frac{x\, dx}{\sqrt{X}} = -\sqrt{X}.$

166. $\int \frac{x^2\, dx}{\sqrt{X}} = -\frac{x}{2}\sqrt{X} + \frac{a^2}{2} \arcsin \frac{x}{a}.$

167. $\int \frac{x^3\, dx}{\sqrt{X}} = \frac{\sqrt{X^3}}{3} - a^2\sqrt{X}.$

168. $\displaystyle\int \frac{dx}{x\sqrt{X}} = -\frac{1}{a}\ln\frac{a+\sqrt{X}}{x}.$

169. $\displaystyle\int \frac{dx}{x^2\sqrt{X}} = -\frac{\sqrt{X}}{a^2 x}.$

170. $\displaystyle\int \frac{dx}{x^3\sqrt{X}} = -\frac{\sqrt{X}}{2a^2x^2} - \frac{1}{2a^3}\ln\frac{a+\sqrt{X}}{x}.$

171. $\displaystyle\int \sqrt{X^3}\,dx = \frac{1}{4}\left(x\sqrt{X^3} + \frac{3a^2x}{2}\sqrt{X} + \frac{3a^4}{2}\arcsin\frac{x}{a}\right).$

172. $\displaystyle\int x\sqrt{X^3}\,dx = -\frac{1}{5}\sqrt{X^5}.$

173. $\displaystyle\int x^2\sqrt{X^3}\,dx = -\frac{x\sqrt{X^5}}{6} + \frac{a^2x\sqrt{X^3}}{24} + \frac{a^4x\sqrt{X}}{16} + \frac{a^6}{16}\arcsin\frac{x}{a}.$

174. $\displaystyle\int x^3\sqrt{X^3}\,dx = \frac{\sqrt{X^7}}{7} - \frac{a^2\sqrt{X^5}}{5}.$

175. $\displaystyle\int \frac{\sqrt{X^3}}{x}\,dx = \frac{\sqrt{X^3}}{3} + a^2\sqrt{X} - a^3\ln\frac{a+\sqrt{X}}{x}.$

176. $\displaystyle\int \frac{\sqrt{X^3}}{x^2}\,dx = -\frac{\sqrt{X^3}}{x} - \frac{3}{2}x\sqrt{X} - \frac{3}{2}a^2\arcsin\frac{x}{a}.$

177. $\displaystyle\int \frac{\sqrt{X^3}}{x^3}\,dx = -\frac{\sqrt{X^3}}{2x^2} - \frac{3\sqrt{X}}{2} + \frac{3a}{2}\ln\frac{a+\sqrt{X}}{x}.$

178. $\displaystyle\int \frac{dx}{\sqrt{X^3}} = \frac{x}{a^2\sqrt{X}}.$

179. $\displaystyle\int \frac{x\,dx}{\sqrt{X^3}} = \frac{1}{\sqrt{X}}.$

180. $\displaystyle\int \frac{x^2\,dx}{\sqrt{X^3}} = \frac{x}{\sqrt{X}} - \arcsin\frac{x}{a}.$

181. $\displaystyle\int \frac{x^3\,dx}{\sqrt{X^3}} = \sqrt{X} + \frac{a^2}{\sqrt{X}}.$

182. $\displaystyle\int \frac{dx}{x\sqrt{X^3}} = \frac{1}{a^2\sqrt{X}} - \frac{1}{a^3}\ln\frac{a+\sqrt{X}}{x}.$

183. $\displaystyle\int \frac{dx}{x^2\sqrt{X^3}} = \frac{1}{a^4}\left(-\frac{\sqrt{X}}{x} + \frac{x}{\sqrt{X}}\right).$

184. $\displaystyle\int \frac{dx}{x^3\sqrt{X^3}} = -\frac{1}{2a^2x^2\sqrt{X}} + \frac{3}{2a^4\sqrt{X}} - \frac{3}{2a^5}\ln\frac{a+\sqrt{X}}{x}.$

21.7.2.6 Integrals with $\sqrt{x^2+a^2}$

Notation: $X = x^2 + a^2$

185. $$\int \sqrt{X}\,dx = \frac{1}{2}\left(x\sqrt{X} + a^2 \operatorname{Arsinh}\frac{x}{a}\right) + C = \frac{1}{2}\left[x\sqrt{X} + a^2 \ln\left(x+\sqrt{X}\right)\right] + C_1 \,.$$

186. $$\int x\sqrt{X}\,dx = \frac{1}{3}\sqrt{X^3}.$$

187. $$\int x^2\sqrt{X}\,dx = \frac{x}{4}\sqrt{X^3} - \frac{a^2}{8}\left(x\sqrt{X} + a^2 \operatorname{Arsinh}\frac{x}{a}\right) + C = \frac{x}{4}\sqrt{X^3} - \frac{a^2}{8}\left[x\sqrt{X} + a^2\ln\left(x+\sqrt{X}\right)\right] + C_1 \,.$$

188. $$\int x^3\sqrt{X}\,dx = \frac{\sqrt{X^5}}{5} - \frac{a^2\sqrt{X^3}}{3}.$$

189. $$\int \frac{\sqrt{X}}{x}\,dx = \sqrt{X} - a\ln\frac{a+\sqrt{X}}{x}.$$

190. $$\int \frac{\sqrt{X}}{x^2}\,dx = -\frac{\sqrt{X}}{x} + \operatorname{Arsinh}\frac{x}{a} + C = -\frac{\sqrt{X}}{x} + \ln\left(x+\sqrt{X}\right) + C_1.$$

191. $$\int \frac{\sqrt{X}}{x^3}\,dx = -\frac{\sqrt{X}}{2x^2} - \frac{1}{2a}\ln\frac{a+\sqrt{X}}{x}.$$

192. $$\int \frac{dx}{\sqrt{X}} = \operatorname{Arsinh}\frac{x}{a} + C = \ln\left(x+\sqrt{X}\right) + C_1.$$

193. $$\int \frac{x\,dx}{\sqrt{X}} = \sqrt{X}.$$

194. $$\int \frac{x^2\,dx}{\sqrt{X}} = \frac{x}{2}\sqrt{X} - \frac{a^2}{2}\operatorname{Arsinh}\frac{x}{a} + C = \frac{x}{2}\sqrt{X} - \frac{a^2}{2}\ln\left(x+\sqrt{X}\right) + C_1.$$

195. $$\int \frac{x^3\,dx}{\sqrt{X}} = \frac{\sqrt{X^3}}{3} - a^2\sqrt{X}.$$

196. $$\int \frac{dx}{x\sqrt{X}} = -\frac{1}{a}\ln\frac{a+\sqrt{X}}{x}.$$

197. $$\int \frac{dx}{x^2\sqrt{X}} = -\frac{\sqrt{X}}{a^2x}.$$

198. $$\int \frac{dx}{x^3\sqrt{X}} = -\frac{\sqrt{X}}{2a^2x^2} + \frac{1}{2a^3}\ln\frac{a+\sqrt{X}}{x}.$$

199. $$\int \sqrt{X^3}\,dx = \frac{1}{4}\left(x\sqrt{X^3} + \frac{3a^2x}{2}\sqrt{X} + \frac{3a^4}{2}\operatorname{Arsinh}\frac{x}{a}\right) + C = \frac{1}{4}\left(x\sqrt{X^3} + \frac{3a^2x}{2}\sqrt{X} + \frac{3a^4}{2}\ln\left(x+\sqrt{X}\right)\right) + C_1.$$

200. $\int x\sqrt{X^3}\,dx = \frac{1}{5}\sqrt{X^5}.$

201. $$\int x^2\sqrt{X^3}\,dx = \frac{x\sqrt{X^5}}{6} - \frac{a^2x\sqrt{X^3}}{24} - \frac{a^4x\sqrt{X}}{16} - \frac{a^6}{16}\operatorname{Arsinh}\frac{x}{a} + C$$
$$= \frac{x\sqrt{X^5}}{6} - \frac{a^2x\sqrt{X^3}}{24} - \frac{a^4x\sqrt{X}}{16} - \frac{a^6}{16}\ln\left(x + \sqrt{X}\right) + C_1\,.$$

202. $$\int x^3\sqrt{X^3}\,dx = \frac{\sqrt{X^7}}{7} - \frac{a^2\sqrt{X^5}}{5}.$$

203. $$\int \frac{\sqrt{X^3}}{x}\,dx = \frac{\sqrt{X^3}}{3} + a^2\sqrt{X} - a^3\ln\frac{a+\sqrt{X}}{x}.$$

204. $$\int \frac{\sqrt{X^3}}{x^2}\,dx = -\frac{\sqrt{X^3}}{x} + \frac{3}{2}x\sqrt{X} + \frac{3}{2}a^2\operatorname{Arsinh}\frac{x}{a} + C$$
$$= -\frac{\sqrt{X^3}}{x} + \frac{3}{2}x\sqrt{X} + \frac{3}{2}a^2\ln\left(x + \sqrt{X}\right) + C_1.$$

205. $$\int \frac{\sqrt{X^3}}{x^3}dx = -\frac{\sqrt{X^3}}{2x^2} + \frac{3}{2}\sqrt{X} - \frac{3}{2}a\ln\left(\frac{a+\sqrt{X}}{x}\right).$$

206. $$\int \frac{dx}{\sqrt{X^3}} = \frac{x}{a^2\sqrt{X}}.$$

207. $$\int \frac{x\,dx}{\sqrt{X^3}} = -\frac{1}{\sqrt{X}}.$$

208. $$\int \frac{x^2\,dx}{\sqrt{X^3}} = -\frac{x}{\sqrt{X}} + \operatorname{Arsinh}\frac{x}{a} + C = -\frac{x}{\sqrt{X}} + \ln\left(x + \sqrt{X}\right) + C_1.$$

209. $$\int \frac{x^3\,dx}{\sqrt{X^3}} = \sqrt{X} + \frac{a^2}{\sqrt{X}}.$$

210. $$\int \frac{dx}{x\sqrt{X^3}} = \frac{1}{a^2\sqrt{X}} - \frac{1}{a^3}\ln\frac{a+\sqrt{X}}{x}.$$

211. $$\int \frac{dx}{x^2\sqrt{X^3}} = -\frac{1}{a^4}\left(\frac{\sqrt{X}}{x} + \frac{x}{\sqrt{X}}\right).$$

212. $$\int \frac{dx}{x^3\sqrt{X^3}} = -\frac{1}{2a^2x^2\sqrt{X}} - \frac{3}{2a^4\sqrt{X}} + \frac{3}{2a^5}\ln\frac{a+\sqrt{X}}{x}.$$

21.7.2.7 Integrals with $\sqrt{x^2 - a^2}$

Notation: $X = x^2 - a^2$

213. $$\int \sqrt{X}\,dx = \frac{1}{2}\left(x\sqrt{X} - a^2\operatorname{Arcosh}\frac{x}{a}\right) + C$$
$$= \frac{1}{2}\left[x\sqrt{X} - a^2\ln\left(x + \sqrt{X}\right)\right] + C_1.$$

214. $$\int x\sqrt{X}\,dx = \frac{1}{3}\sqrt{X^3}.$$

215. $$\int x^2\sqrt{X}\,dx = \frac{x}{4}\sqrt{X^3} + \frac{a^2}{8}\left(x\sqrt{X} - a^2 \operatorname{Arcosh}\frac{x}{a}\right) + C$$
$$= \frac{x}{4}\sqrt{X^3} + \frac{a^2}{8}\left[x\sqrt{X} - a^2 \ln\left(x + \sqrt{X}\right)\right] + C_1.$$

216. $$\int x^3\sqrt{X}\,dx = \frac{\sqrt{X^5}}{5} + \frac{a^2\sqrt{X^3}}{3}.$$

217. $$\int \frac{\sqrt{X}}{x}\,dx = \sqrt{X} - a\arccos\frac{a}{x}.$$

218. $$\int \frac{\sqrt{X}}{x^2}\,dx = -\frac{\sqrt{X}}{x} + \operatorname{Arcosh}\frac{x}{a} + C = -\frac{\sqrt{X}}{x} + \ln\left(x + \sqrt{X}\right) + C_1.$$

219. $$\int \frac{\sqrt{X}}{x^3}\,dx = -\frac{\sqrt{X}}{2x^2} + \frac{1}{2a}\arccos\frac{a}{x}.$$

220. $$\int \frac{dx}{\sqrt{X}} = \operatorname{Arcosh}\frac{x}{a} + C = \ln\left(x + \sqrt{X}\right) + C_1.$$

221. $$\int \frac{x\,dx}{\sqrt{X}} = \sqrt{X}.$$

222. $$\int \frac{x^2\,dx}{\sqrt{X}} = \frac{x}{2}\sqrt{X} + \frac{a^2}{2}\operatorname{Arcosh}\frac{x}{a} + C = \frac{x}{2}\sqrt{X} + \frac{a^2}{2}\ln\left(x + \sqrt{X}\right) + C_1.$$

223. $$\int \frac{x^3\,dx}{\sqrt{X}} = \frac{\sqrt{X^3}}{3} + a^2\sqrt{X}.$$

224. $$\int \frac{dx}{x\sqrt{X}} = \frac{1}{a}\arccos\frac{a}{x}.$$

225. $$\int \frac{dx}{x^2\sqrt{X}} = \frac{\sqrt{X}}{a^2 x}.$$

226. $$\int \frac{dx}{x^3\sqrt{X}} = \frac{\sqrt{X}}{2a^2x^2} + \frac{1}{2a^3}\arccos\frac{a}{x}.$$

227. $$\int \sqrt{X^3}\,dx = \frac{1}{4}\left(x\sqrt{X^3} - \frac{3a^2x}{2}\sqrt{X} + \frac{3a^4}{2}\operatorname{Arcosh}\frac{x}{a}\right) + C$$
$$= \frac{1}{4}\left(x\sqrt{X^3} - \frac{3a^2x}{2}\sqrt{X} + \frac{3a^4}{2}\ln\left(x + \sqrt{X}\right)\right) + C_1.$$

228. $$\int x\sqrt{X^3}\,dx = \frac{1}{5}\sqrt{X^5}.$$

229. $$\int x^2\sqrt{X^3}\,dx = \frac{x\sqrt{X^5}}{6} + \frac{a^2x\sqrt{X^3}}{24} - \frac{a^4x\sqrt{X}}{16} + \frac{a^6}{16}\operatorname{Arcosh}\frac{x}{a} + C$$
$$= \frac{x\sqrt{X^5}}{6} + \frac{a^2x\sqrt{X^3}}{24} - \frac{a^4x\sqrt{X}}{16} + \frac{a^6}{16}\ln\left(x + \sqrt{X}\right) + C_1.$$

230. $$\int x^3\sqrt{X^3}\,dx = \frac{\sqrt{X^7}}{7} + \frac{a^2\sqrt{X^5}}{5}.$$

231. $\displaystyle\int \frac{\sqrt{X^3}}{x}\,dx = \frac{\sqrt{X^3}}{3} - a^2\sqrt{X} + a^3 \arccos\frac{a}{x}.$

232. $\displaystyle\int \frac{\sqrt{X^3}}{x^2}\,dx = -\frac{\sqrt{X^3}}{2} + \frac{3}{2}x\sqrt{X} - \frac{3}{2}a^2 \operatorname{Arcosh}\frac{x}{a} + C$
$\displaystyle = -\frac{\sqrt{X^3}}{2} + \frac{3}{2}x\sqrt{X} - \frac{3}{2}a^2 \ln\left(x+\sqrt{X}\right) + C_1.$

233. $\displaystyle\int \frac{\sqrt{X^3}}{x^3}\,dx = -\frac{\sqrt{X^3}}{2x^2} + \frac{3\sqrt{X}}{2} - \frac{3}{2}a \arccos\frac{a}{x}.$

234. $\displaystyle\int \frac{dx}{\sqrt{X^3}} = -\frac{x}{a^2\sqrt{X}}.$

235. $\displaystyle\int \frac{x\,dx}{\sqrt{X^3}} = -\frac{1}{\sqrt{X}}.$

236. $\displaystyle\int \frac{x^2\,dx}{\sqrt{X^3}} = -\frac{x}{\sqrt{X}} + \operatorname{Arcosh}\frac{x}{a} + C = -\frac{x}{\sqrt{X}} + \ln\left(x+\sqrt{X}\right) + C_1.$

237. $\displaystyle\int \frac{x^3\,dx}{\sqrt{X^3}} = \sqrt{X} - \frac{a^2}{\sqrt{X}}.$

238. $\displaystyle\int \frac{dx}{x\sqrt{X^3}} = -\frac{1}{a^2\sqrt{X}} - \frac{1}{a^3}\arccos\frac{a}{x}.$

239. $\displaystyle\int \frac{dx}{x^2\sqrt{X^3}} = -\frac{1}{a^4}\left(\frac{\sqrt{X}}{x} + \frac{x}{\sqrt{X}}\right).$

240. $\displaystyle\int \frac{dx}{x^3\sqrt{X^3}} = \frac{1}{2a^2x^2\sqrt{X}} - \frac{3}{2a^4\sqrt{X}} - \frac{3}{2a^5}\arccos\frac{a}{x}.$

21.7.2.8 Integrals with $\sqrt{ax^2+bx+c}$

Notation: $X = ax^2+bx+c\,,\ \Delta = 4ac-b^2\,,\ k = \dfrac{4a}{\Delta}$

241. $$\int \frac{dx}{\sqrt{X}} = \begin{cases} \dfrac{1}{\sqrt{a}}\ln\left(2\sqrt{aX}+2ax+b\right)+C & \text{for } a>0\,, \\ \dfrac{1}{\sqrt{a}}\operatorname{Arsinh}\dfrac{2ax+b}{\sqrt{\Delta}}+C_1 & \text{for } a>0\,,\ \Delta>0\,, \\ \dfrac{1}{\sqrt{a}}\ln(2ax+b) & \text{for } a>0\,,\ \Delta=0\,, \\ -\dfrac{1}{\sqrt{-a}}\arcsin\dfrac{2ax+b}{\sqrt{-\Delta}} & \text{for } a<0\,,\ \Delta<0\,. \end{cases}$$

242. $\displaystyle\int \frac{dx}{X\sqrt{X}} = \frac{2(2ax+b)}{\Delta\sqrt{X}}.$

243. $\displaystyle\int \frac{dx}{X^2\sqrt{X}} = \frac{2(2ax+b)}{3\Delta\sqrt{X}}\left(\frac{1}{X}+2k\right).$

244. $\displaystyle\int \frac{dx}{X^{(2n+1)/2}} = \frac{2(2ax+b)}{(2n-1)\Delta X^{(2n-1)/2}} + \frac{2k(n-1)}{2n-1}\int \frac{dx}{X^{(2n-1)/2}}.$

245. $\int \sqrt{X}\,dx = \frac{(2ax+b)\sqrt{X}}{4a} + \frac{1}{2k}\int \frac{dx}{\sqrt{X}}$ (see No. 241).

246. $\int X\sqrt{X}\,dx = \frac{(2ax+b)\sqrt{X}}{8a}\left(X + \frac{3}{2k}\right) + \frac{3}{8k^2}\int \frac{dx}{\sqrt{X}}$ (see No. 241).

247. $\int X^2\sqrt{X}\,dx = \frac{(2ax+b)\sqrt{X}}{12a}\left(X^2 + \frac{5X}{4k} + \frac{15}{8k^2}\right) + \frac{5}{16k^3}\int \frac{dx}{\sqrt{X}}$ (see No. 241).

248. $\int X^{(2n+1)/2}\,dx = \frac{(2ax+b)X^{(2n+1)/2}}{4a(n+1)} + \frac{2n+1}{2k(n+1)}\int X^{(2n-1)/2}\,dx.$

249. $\int \frac{x\,dx}{\sqrt{X}} = \frac{\sqrt{X}}{a} - \frac{b}{2a}\int \frac{dx}{\sqrt{X}}$ (see No. 241).

250. $\int \frac{x\,dx}{X\sqrt{X}} = -\frac{2(bx+2c)}{\Delta\sqrt{X}}.$

251. $\int \frac{x\,dx}{X^{(2n+1)/2}} = -\frac{1}{(2n-1)aX^{(2n-1)/2}} - \frac{b}{2a}\int \frac{dx}{X^{(2n+1)/2}}$ (see No. 244).

252. $\int \frac{x^2\,dx}{\sqrt{X}} = \left(\frac{x}{2a} - \frac{3b}{4a^2}\right)\sqrt{X} + \frac{3b^2-4ac}{8a^2}\int \frac{dx}{\sqrt{X}}$ (see No. 241).

253. $\int \frac{x^2\,dx}{X\sqrt{X}} = \frac{(2b^2-4ac)x+2bc}{a\Delta\sqrt{X}} + \frac{1}{a}\int \frac{dx}{\sqrt{X}}$ (see No. 241).

254. $\int x\sqrt{X}\,dx = \frac{X\sqrt{X}}{3a} - \frac{b(2ax+b)}{8a^2}\sqrt{X} - \frac{b}{4ak}\int \frac{dx}{\sqrt{X}}$ (see No. 241).

255. $\int xX\sqrt{X}\,dx = \frac{X^2\sqrt{X}}{5a} - \frac{b}{2a}\int X\sqrt{X}\,dx$ (see No. 246).

256. $\int xX^{(2n+1)/2}\,dx = \frac{X^{(2n+3)/2}}{(2n+3)a} - \frac{b}{2a}\int X^{(2n+1)/2}\,dx$ (see No. 248).

257. $\int x^2\sqrt{X}\,dx = \left(x - \frac{5b}{6a}\right)\frac{X\sqrt{X}}{4a} + \frac{5b^2-4ac}{16a^2}\int \sqrt{X}\,dx$ (see No. 245).

258. $$\int \frac{dx}{x\sqrt{X}} = \begin{cases} -\frac{1}{\sqrt{c}}\ln\left(\frac{2\sqrt{cX}}{x} + \frac{2c}{x} + b\right) + C & \text{for } c>0\,, \\ -\frac{1}{\sqrt{c}}\operatorname{Arsinh}\frac{bx+2c}{x\sqrt{\Delta}} + C_1 & \text{for } c>0\,,\ \Delta>0\,, \\ -\frac{1}{\sqrt{c}}\ln\frac{bx+2c}{x} & \text{for } c>0\,,\ \Delta=0\,, \\ \frac{1}{\sqrt{-c}}\arcsin\frac{bx+2c}{x\sqrt{-\Delta}} & \text{for } c<\,,\ \Delta<0\,. \end{cases}$$

259. $\int \frac{dx}{x^2\sqrt{X}} = -\frac{\sqrt{X}}{cx} - \frac{b}{2c}\int \frac{dx}{x\sqrt{X}}$ (see No. 258).

260. $\int \frac{\sqrt{X}\,dx}{x} = \sqrt{X} + \frac{b}{2}\int \frac{dx}{\sqrt{X}} + c\int \frac{dx}{x\sqrt{X}}$ (see No. 241 and 258).

261. $\displaystyle\int \frac{\sqrt{X}\,dx}{x^2} = -\frac{\sqrt{X}}{x} + a\int \frac{dx}{\sqrt{X}} + \frac{b}{2}\int \frac{dx}{x\sqrt{X}}$ (see No. 241 and 258).

262. $\displaystyle\int \frac{X^{(2n+1)/2}}{x} dx = \frac{X^{(2n+1)/2}}{2n+1} + \frac{b}{2}\int X^{(2n-1)/2}\,dx + c\int \frac{X^{(2n-1)/2}}{x} dx$ (see No. 248 and 260).

263. $\displaystyle\int \frac{dx}{x\sqrt{ax^2+bx}} = -\frac{2}{bx}\sqrt{ax^2+bx}.$

264. $\displaystyle\int \frac{dx}{\sqrt{2ax-x^2}} = \arcsin\frac{x-a}{a}.$

265. $\displaystyle\int \frac{x\,dx}{\sqrt{2ax-x^2}} = -\sqrt{2ax-x^2} + a\arcsin\frac{x-a}{a}.$

266. $\displaystyle\int \sqrt{2ax-x^2}\,dx = \frac{x-a}{2}\sqrt{2ax-x^2} + \frac{a^2}{2}\arcsin\frac{x-a}{a}.$

267. $\displaystyle\int \frac{dx}{(ax^2+b)\sqrt{fx^2+g}} = \frac{1}{\sqrt{b}\sqrt{ag-bf}}\arctan\frac{x\sqrt{ag-bf}}{\sqrt{b}\sqrt{fx^2+g}}$ $(ag-bf>0)\,,$

$\displaystyle = \frac{1}{2\sqrt{b}\sqrt{bf-ag}}\ln\frac{\sqrt{b}\sqrt{fx^2+g}+x\sqrt{bf-ag}}{\sqrt{b}\sqrt{fx^2+g}-x\sqrt{bf-ag}}$ $(ag-bf<0)\,.$

21.7.2.9 Integrals with other Irrational Expressions

268. $\displaystyle\int \sqrt[n]{ax+b}\,dx = \frac{n(ax+b)}{(n+1)a}\sqrt[n]{ax+b}.$

269. $\displaystyle\int \frac{dx}{\sqrt[n]{ax+b}} = \frac{n(ax+b)}{(n-1)a}\frac{1}{\sqrt[n]{ax+b}}.$

270. $\displaystyle\int \frac{dx}{x\sqrt{x^n+a^2}} = -\frac{2}{na}\ln\frac{a+\sqrt{x^n+a^2}}{\sqrt{x^n}}.$

271. $\displaystyle\int \frac{dx}{x\sqrt{x^n-a^2}} = \frac{2}{na}\arccos\frac{a}{\sqrt{x^n}}.$

272. $\displaystyle\int \frac{\sqrt{x}\,dx}{\sqrt{a^3-x^3}} = \frac{2}{3}\arcsin\sqrt{\left(\frac{x}{a}\right)^3}.$

21.7.2.10 Recursion Formulas for an Integral with Binomial Differential

273. $\displaystyle\int x^m(ax^n+b)^p\,dx$

$$= \frac{1}{m+np+1}\left[x^{m+1}(ax^n+b)^p + npb\int x^m(ax^n+b)^{p-1}\,dx\right],$$

$$= \frac{1}{bn(p+1)}\left[-x^{m+1}(ax^n+b)^{p+1} + (m+n+np+1)\int x^m(ax^n+b)^{p+1}\,dx\right],$$

$$= \frac{1}{(m+1)b}\left[x^{m+1}(ax^n+b)^{p+1} - a(m+n+np+1)\int x^{m+n}(ax^n+b)^p\,dx\right],$$

$$= \frac{1}{a(m+np+1)}\left[x^{m-n+1}(ax^n+b)^{p+1} - (m-n+1)b\int x^{m-n}(ax^n+b)^p\,dx\right].$$

21.7.3 Integrals of Trigonometric Functions

Integrals of functions also containing $\sin x$ and $\cos x$ together with hyperbolic and exponential functions are in the table of integrals of other transcendental functions (see 21.7.4, p. 1092).

21.7.3.1 Integrals with Sine Function

274. $\int \sin ax\,dx = -\frac{1}{a}\cos ax.$

275. $\int \sin^2 ax\,dx = \frac{1}{2}x - \frac{1}{4a}\sin 2ax.$

276. $\int \sin^3 ax\,dx = -\frac{1}{a}\cos ax + \frac{1}{3a}\cos^3 ax.$

277. $\int \sin^4 ax\,dx = \frac{3}{8}x - \frac{1}{4a}\sin 2ax + \frac{1}{32a}\sin 4ax.$

278. $\int \sin^n ax\,dx = -\frac{\sin^{n-1} ax\cos ax}{na} + \frac{n-1}{n}\int \sin^{n-2} ax\,dx$ (n integer numbers, > 0).

279. $\int x\sin ax\,dx = \frac{\sin ax}{a^2} - \frac{x\cos ax}{a}.$

280. $\int x^2\sin ax\,dx = \frac{2x}{a^2}\sin ax - \left(\frac{x^2}{a} - \frac{2}{a^3}\right)\cos ax.$

281. $\int x^3\sin ax\,dx = \left(\frac{3x^2}{a^2} - \frac{6}{a^4}\right)\sin ax - \left(\frac{x^3}{a} - \frac{6x}{a^3}\right)\cos ax.$

282. $\int x^n\sin ax\,dx = -\frac{x^n}{a}\cos ax + \frac{n}{a}\int x^{n-1}\cos ax\,dx$ ($n > 0$).

283. $\int \frac{\sin ax}{x}dx = ax - \frac{(ax)^3}{3\cdot 3!} + \frac{(ax)^5}{5\cdot 5!} - \frac{(ax)^7}{7\cdot 7!} + \cdots.$

The definite integral $\int\limits_0^x \frac{\sin t}{t}\,dt$ is called the sine integral (see 8.2.5, **1.**, p. 513) and it is denoted $\mathrm{si}(x)$. For the calculation of the integral see 14.4.3.2, **2.**, p. 756. The power series expansion is
$\mathrm{si}(x) = x - \frac{x^3}{3\cdot 3!} + \frac{x^5}{5\cdot 5!} - \frac{x^7}{7\cdot 7!} + \cdots$; see 8.2.5, **1.**, p. 513.

284. $\int \frac{\sin ax}{x^2}dx = -\frac{\sin ax}{x} + a\int\frac{\cos ax\,dx}{x}$ (see No. 322).

285. $\int \frac{\sin ax}{x^n}dx = -\frac{1}{n-1}\frac{\sin ax}{x^{n-1}} + \frac{a}{n-1}\int\frac{\cos ax}{x^{n-1}}dx$ (see No. 324).

286. $\int \frac{dx}{\sin ax} = \int \operatorname{cosec} ax\,dx = \frac{1}{a}\ln\tan\frac{ax}{2} = \frac{1}{a}\ln(\operatorname{cosec} ax - \cot ax).$

287. $\int \frac{dx}{\sin^2 ax} = -\frac{1}{a}\cot ax.$

288. $\displaystyle\int \frac{dx}{\sin^3 ax} = -\frac{\cos ax}{2a\sin^2 ax} + \frac{1}{2a}\ln\tan\frac{ax}{2}.$

289. $\displaystyle\int \frac{dx}{\sin^n ax} = -\frac{1}{a(n-1)}\frac{\cos ax}{\sin^{n-1} ax} + \frac{n-2}{n-1}\int\frac{dx}{\sin^{n-2} ax} \qquad (n > 1).$

290. $\displaystyle\int \frac{x\,dx}{\sin ax} = \frac{1}{a^2}\left(ax + \frac{(ax)^3}{3\cdot 3!} + \frac{7(ax)^5}{3\cdot 5\cdot 5!} + \frac{31(ax)^7}{3\cdot 7\cdot 7!} + \frac{127(ax)^9}{3\cdot 5\cdot 9!} + \cdots + \frac{2(2^{2n-1}-1)}{(2n+1)!}B_n(ax)^{2n+1} + \cdots\right)$

B_n denote the Bernoulli numbers (see 7.2.4.2, p. 465).

291. $\displaystyle\int \frac{x\,dx}{\sin^2 ax} = -\frac{x}{a}\cot ax + \frac{1}{a^2}\ln\sin ax.$

292. $\displaystyle\int \frac{x\,dx}{\sin^n ax} = -\frac{x\cos ax}{(n-1)a\sin^{n-1} ax} - \frac{1}{(n-1)(n-2)a^2\sin^{n-2} ax} + \frac{n-2}{n-1}\int\frac{x\,dx}{\sin^{n-2} ax} \qquad (n > 2).$

293. $\displaystyle\int \frac{dx}{1+\sin ax} = -\frac{1}{a}\tan\left(\frac{\pi}{4} - \frac{ax}{2}\right).$

294. $\displaystyle\int \frac{dx}{1-\sin ax} = \frac{1}{a}\tan\left(\frac{\pi}{4} + \frac{ax}{2}\right).$

295. $\displaystyle\int \frac{x\,dx}{1+\sin ax} = -\frac{x}{a}\tan\left(\frac{\pi}{4} - \frac{ax}{2}\right) + \frac{2}{a^2}\ln\cos\left(\frac{\pi}{4} - \frac{ax}{2}\right).$

296. $\displaystyle\int \frac{x\,dx}{1-\sin ax} = \frac{x}{a}\cot\left(\frac{\pi}{4} - \frac{ax}{2}\right) + \frac{2}{a^2}\ln\sin\left(\frac{\pi}{4} - \frac{ax}{2}\right).$

297. $\displaystyle\int \frac{\sin ax\,dx}{1\pm\sin ax} = \pm x + \frac{1}{a}\tan\left(\frac{\pi}{4} \mp \frac{ax}{2}\right).$

298. $\displaystyle\int \frac{dx}{\sin ax(1\pm\sin ax)} = \frac{1}{a}\tan\left(\frac{\pi}{4} \mp \frac{ax}{2}\right) + \frac{1}{a}\ln\tan\frac{ax}{2}.$

299. $\displaystyle\int \frac{dx}{(1+\sin ax)^2} = -\frac{1}{2a}\tan\left(\frac{\pi}{4} - \frac{ax}{2}\right) - \frac{1}{6a}\tan^3\left(\frac{\pi}{4} - \frac{ax}{2}\right).$

300. $\displaystyle\int \frac{dx}{(1-\sin ax)^2} = \frac{1}{2a}\cot\left(\frac{\pi}{4} - \frac{ax}{2}\right) + \frac{1}{6a}\cot^3\left(\frac{\pi}{4} - \frac{ax}{2}\right).$

301. $\displaystyle\int \frac{\sin ax\,dx}{(1+\sin ax)^2} = -\frac{1}{2a}\tan\left(\frac{\pi}{4} - \frac{ax}{2}\right) + \frac{1}{6a}\tan^3\left(\frac{\pi}{4} - \frac{ax}{2}\right).$

302. $\displaystyle\int \frac{\sin ax\,dx}{(1-\sin ax)^2} = -\frac{1}{2a}\cot\left(\frac{\pi}{4} - \frac{ax}{2}\right) + \frac{1}{6a}\cot^3\left(\frac{\pi}{4} - \frac{ax}{2}\right).$

303. $\displaystyle\int \frac{dx}{1+\sin^2 ax} = \frac{1}{2\sqrt{2}a}\arcsin\left(\frac{3\sin^2 ax - 1}{\sin^2 ax + 1}\right).$

304. $\displaystyle\int \frac{dx}{1-\sin^2 ax} = \int\frac{dx}{\cos^2 ax} = \frac{1}{a}\tan ax.$

305. $\int \sin ax \sin bx \, dx = \frac{\sin(a-b)x}{2(a-b)} - \frac{\sin(a+b)x}{2(a+b)}$ $\quad (|a| \neq |b|;$ $\quad$ for $|a| = |b|$ see No.275).

306. $\int \frac{dx}{b + c \sin ax} = \frac{2}{a\sqrt{b^2 - c^2}} \arctan \frac{b \tan ax/2 + c}{\sqrt{b^2 - c^2}}$ $\quad$ for $b^2 > c^2$),

$$= \frac{1}{a\sqrt{c^2 - b^2}} \ln \frac{b \tan ax/2 + c - \sqrt{c^2 - b^2}}{b \tan ax/2 + c + \sqrt{c^2 - b^2}}$$ for $b^2 < c^2$).

307. $\int \frac{\sin ax \, dx}{b + c \sin ax} = \frac{x}{c} - \frac{b}{c} \int \frac{dx}{b + c \sin ax}$ $\quad$ (see No. 306).

308. $\int \frac{dx}{\sin ax (b + c \sin ax)} = \frac{1}{ab} \ln \tan \frac{ax}{2} - \frac{c}{b} \int \frac{dx}{b + c \sin ax}$ $\quad$ (see No. 306).

309. $\int \frac{dx}{(b + c \sin ax)^2} = \frac{c \cos ax}{a(b^2 - c^2)(b + c \sin ax)} + \frac{b}{b^2 - c^2} \int \frac{dx}{b + c \sin ax}$ $\quad$ (see No. 306).

310. $\int \frac{\sin ax \, dx}{(b + c \sin ax)^2} = \frac{b \cos ax}{a(c^2 - b^2)(b + c \sin ax)} + \frac{c}{c^2 - b^2} \int \frac{dx}{b + c \sin ax}$ $\quad$ (see No. 306).

311. $\int \frac{dx}{b^2 + c^2 \sin^2 ax} = \frac{1}{ab\sqrt{b^2 + c^2}} \arctan \frac{\sqrt{b^2 + c^2} \tan ax}{b}$ $\quad (b > 0)$.

312. $\int \frac{dx}{b^2 - c^2 \sin^2 ax} = \frac{1}{ab\sqrt{b^2 - c^2}} \arctan \frac{\sqrt{b^2 - c^2} \tan ax}{b}$ $\quad (b^2 > c^2,\ b > 0)$,

$$= \frac{1}{2ab\sqrt{c^2 - b^2}} \ln \frac{\sqrt{c^2 - b^2} \tan ax + b}{\sqrt{c^2 - b^2} \tan ax - b}$$ $(c^2 > b^2,\ b > 0)$.

21.7.3.2 Integrals with Cosine Function

313. $\int \cos ax \, dx = \frac{1}{a} \sin ax.$

314. $\int \cos^2 ax \, dx = \frac{1}{2} x + \frac{1}{4a} \sin 2ax.$

315. $\int \cos^3 ax \, dx = \frac{1}{a} \sin ax - \frac{1}{3a} \sin^3 ax.$

316. $\int \cos^4 ax \, dx = \frac{3}{8} x + \frac{1}{4a} \sin 2ax + \frac{1}{32a} \sin 4ax.$

317. $\int \cos^n ax \, dx = \frac{\cos^{n-1} ax \sin ax}{na} + \frac{n-1}{n} \int \cos^{n-2} ax \, dx.$

318. $\int x \cos ax \, dx = \frac{\cos ax}{a^2} + \frac{x \sin ax}{a}.$

319. $\int x^2 \cos ax \, dx = \frac{2x}{a^2} \cos ax + \left(\frac{x^2}{a} - \frac{2}{a^3} \right) \sin ax.$

320. $\int x^3 \cos ax \, dx = \left(\frac{3x^2}{a^2} - \frac{6}{a^4} \right) \cos ax + \left(\frac{x^3}{a} - \frac{6x}{a^3} \right) \sin ax.$

321. $\int x^n \cos ax \, dx = \frac{x^n \sin ax}{a} - \frac{n}{a} \int x^{n-1} \sin ax \, dx.$

322. $\displaystyle\int \frac{\cos ax}{x}dx = \ln(ax) - \frac{(ax)^2}{2\cdot 2!} + \frac{(ax)^4}{4\cdot 4!} - \frac{(ax)^6}{6\cdot 6!} + \cdots$

The definite integral $-\displaystyle\int_x^\infty \frac{\cos t}{t}\,dt$ is called the cosine integral (see 14.4.3.2, p. 756) and it is denoted by $\mathrm{Ci}(x)$. The power series expansion is $\mathrm{Ci}(x) = C + \ln x - \dfrac{x^2}{2\cdot 2!} + \dfrac{x^4}{4\cdot 4!} - \dfrac{x^6}{6\cdot 6!} + \cdots$ see 8.2.5, **2.**, p. 513; C denotes the Euler constant (see 8.2.5, **2.**, p. 513).

323. $\displaystyle\int \frac{\cos ax}{x^2}dx = -\frac{\cos ax}{x} - a\int \frac{\sin ax\,dx}{x}$ (see No. 283).

324. $\displaystyle\int \frac{\cos ax}{x^n}dx = -\frac{\cos ax}{(n-1)x^{n-1}} - \frac{a}{n-1}\int \frac{\sin ax\,dx}{x^{n-1}} \quad (n \neq 1)$ (see No. 285).

325. $\displaystyle\int \frac{dx}{\cos ax} = \frac{1}{a}\,\mathrm{Artanh}(\sin ax) = \frac{1}{a}\ln\tan\left(\frac{ax}{2} + \frac{\pi}{4}\right) = \frac{1}{a}\ln(\sec ax + \tan ax).$

326. $\displaystyle\int \frac{dx}{\cos^2 ax} = \frac{1}{a}\tan ax.$

327. $\displaystyle\int \frac{dx}{\cos^3 ax} = \frac{\sin ax}{2a\cos^2 ax} + \frac{1}{2a}\ln\tan\left(\frac{\pi}{4} + \frac{ax}{2}\right).$

328. $\displaystyle\int \frac{dx}{\cos^n ax} = \frac{1}{a(n-1)}\frac{\sin ax}{\cos^{n-1} ax} + \frac{n-2}{n-1}\int \frac{dx}{\cos^{n-2} ax} \quad (n > 1).$

329. $\displaystyle\int \frac{x\,dx}{\cos ax} = \frac{1}{a^2}\left(\frac{(ax)^2}{2} + \frac{(ax)^4}{4\cdot 2!} + \frac{5(ax)^6}{6\cdot 4!} + \frac{61(ax)^8}{8\cdot 6!} + \frac{1385(ax)^{10}}{10\cdot 8!} + \cdots + \frac{E_n(ax)^{2n+2}}{(2n+2)(2n!)} + \cdots\right)$

E_n denote the Euler numbers (see 7.2, p. 466).

330. $\displaystyle\int \frac{x\,dx}{\cos^2 ax} = \frac{x}{a}\tan ax + \frac{1}{a^2}\ln\cos ax.$

331. $\displaystyle\int \frac{x\,dx}{\cos^n ax} = \frac{x\sin ax}{(n-1)a\cos^{n-1} ax} - \frac{1}{(n-1)(n-2)a^2\cos^{n-2} ax} + \frac{n-2}{n-1}\int \frac{x\,dx}{\cos^{n-2} ax} \quad (n > 2).$

332. $\displaystyle\int \frac{dx}{1+\cos ax} = \frac{1}{a}\tan\frac{ax}{2}.$

333. $\displaystyle\int \frac{dx}{1-\cos ax} = -\frac{1}{a}\cot\frac{ax}{2}.$

334. $\displaystyle\int \frac{x\,dx}{1+\cos ax} = \frac{x}{a}\tan\frac{ax}{2} + \frac{2}{a^2}\ln\cos\frac{ax}{2}.$

335. $\displaystyle\int \frac{x\,dx}{1-\cos ax} = -\frac{x}{a}\cot\frac{ax}{2} + \frac{2}{a^2}\ln\sin\frac{ax}{2}.$

336. $\displaystyle\int \frac{\cos ax\,dx}{1+\cos ax} = x - \frac{1}{a}\tan\frac{ax}{2}.$

337. $\displaystyle\int \frac{\cos ax\,dx}{1-\cos ax} = -x - \frac{1}{a}\cot\frac{ax}{2}.$

338. $\displaystyle\int \frac{dx}{\cos ax(1+\cos ax)} = \frac{1}{a}\ln\tan\left(\frac{\pi}{4} + \frac{ax}{2}\right) - \frac{1}{a}\tan\frac{ax}{2}.$

339. $\int \frac{dx}{\cos ax(1-\cos ax)} = \frac{1}{a}\ln\tan\left(\frac{\pi}{4}+\frac{ax}{2}\right) - \frac{1}{a}\cot\frac{ax}{2}.$

340. $\int \frac{dx}{(1+\cos ax)^2} = \frac{1}{2a}\tan\frac{ax}{2} + \frac{1}{6a}\tan^3\frac{ax}{2}.$

341. $\int \frac{dx}{(1-\cos ax)^2} = -\frac{1}{2a}\cot\frac{ax}{2} - \frac{1}{6a}\cot^3\frac{ax}{2}.$

342. $\int \frac{\cos ax\,dx}{(1+\cos ax)^2} = \frac{1}{2a}\tan\frac{ax}{2} - \frac{1}{6a}\tan^3\frac{ax}{2}.$

343. $\int \frac{\cos ax\,dx}{(1-\cos ax)^2} = \frac{1}{2a}\cot\frac{ax}{2} - \frac{1}{6a}\cot^3\frac{ax}{2}.$

344. $\int \frac{dx}{1+\cos^2 ax} = \frac{1}{2\sqrt{2}a}\arcsin\left(\frac{1-3\cos^2 ax}{1+\cos^2 ax}\right).$

345. $\int \frac{dx}{1-\cos^2 ax} = \int \frac{dx}{\sin^2 ax} = -\frac{1}{a}\cot ax.$

346. $\int \cos ax\cos bx\,dx = \frac{\sin(a-b)x}{2(a-b)} + \frac{\sin(a+b)x}{2(a+b)}$ $\quad (|a|\neq|b|)$; (for $|a|=|b|$ see No. 314).

347. $\int \frac{dx}{b+c\cos ax} = \frac{2}{a\sqrt{b^2-c^2}}\arctan\frac{(b-c)\tan ax/2}{\sqrt{b^2-c^2}}$ (for $b^2>c^2$)

$= \frac{1}{a\sqrt{c^2-b^2}}\ln\frac{(c-b)\tan ax/2+\sqrt{c^2-b^2}}{(c-b)\tan ax/2-\sqrt{c^2-b^2}}$ (for $b^2<c^2$).

348. $\int \frac{\cos ax\,dx}{b+c\cos ax} = \frac{x}{c} - \frac{b}{c}\int\frac{dx}{b+c\cos ax}$ (see No. 347).

349. $\int \frac{dx}{\cos ax(b+c\cos ax)} = \frac{1}{ab}\ln\tan\left(\frac{ax}{2}+\frac{\pi}{4}\right) - \frac{c}{b}\int\frac{dx}{b+c\cos ax}$ (see No. 347).

350. $\int \frac{dx}{(b+c\cos ax)^2} = \frac{c\sin ax}{a(c^2-b^2)(b+c\cos ax)} - \frac{b}{c^2-b^2}\int\frac{dx}{b+c\cos ax}$ (see No. 347).

351. $\int \frac{\cos ax\,dx}{(b+c\cos ax)^2} = \frac{b\sin ax}{a(b^2-c^2)(b+c\cos ax)} - \frac{c}{b^2-c^2}\int\frac{dx}{b+c\cos ax}$ (see No. 347).

352. $\int \frac{dx}{b^2+c^2\cos^2 ax} = \frac{1}{ab\sqrt{b^2+c^2}}\arctan\frac{b\tan ax}{\sqrt{b^2+c^2}}$ $\quad (b>0)$.

353. $\int \frac{dx}{b^2-c^2\cos^2 ax} = \frac{1}{ab\sqrt{b^2-c^2}}\arctan\frac{b\tan ax}{\sqrt{b^2-c^2}}$ $\quad (b^2>c^2,\ b>0)$,

$= \frac{1}{2ab\sqrt{c^2-b^2}}\ln\frac{b\tan ax-\sqrt{c^2-b^2}}{b\tan ax+\sqrt{c^2-b^2}}$ $\quad (c^2>b^2,\ b>0)$.

21.7.3.3 Integrals with Sine and Cosine Function

354. $\int \sin ax\cos ax\,dx = \frac{1}{2a}\sin^2 ax.$

355. $\displaystyle\int \sin^2 ax \cos^2 ax\, dx = \frac{x}{8} - \frac{\sin 4ax}{32a}.$

356. $\displaystyle\int \sin^n ax \cos ax\, dx = \frac{1}{a(n+1)} \sin^{n+1} ax \qquad (n \neq -1).$

357. $\displaystyle\int \sin ax \cos^n ax\, dx = -\frac{1}{a(n+1)} \cos^{n+1} ax \qquad (n \neq -1).$

358. $$\int \sin^n ax \cos^m ax\, dx = -\frac{\sin^{n-1} ax \cos^{m+1} ax}{a(n+m)} + \frac{n-1}{n+m}\int \sin^{n-2} ax \cos^m ax\, dx$$

(lowering the exponent n; m and $n > 0$),

$$= \frac{\sin^{n+1} ax \cos^{m-1} ax}{a(n+m)} + \frac{m-1}{n+m}\int \sin^n ax \cos^{m-2} ax\, dx$$

(lowering the exponent m; m and $n > 0$).

359. $\displaystyle\int \frac{dx}{\sin ax \cos ax} = \frac{1}{a} \ln \tan ax.$

360. $\displaystyle\int \frac{dx}{\sin^2 ax \cos ax} = \frac{1}{a}\left[\ln \tan\left(\frac{\pi}{4} + \frac{ax}{2}\right) - \frac{1}{\sin ax}\right].$

361. $\displaystyle\int \frac{dx}{\sin ax \cos^2 ax} = \frac{1}{a}\left(\ln \tan \frac{ax}{2} + \frac{1}{\cos ax}\right).$

362. $\displaystyle\int \frac{dx}{\sin^3 ax \cos ax} = \frac{1}{a}\left(\ln \tan ax - \frac{1}{2\sin^2 ax}\right).$

363. $\displaystyle\int \frac{dx}{\sin ax \cos^3 ax} = \frac{1}{a}\left(\ln \tan ax + \frac{1}{2\cos^2 ax}\right).$

364. $\displaystyle\int \frac{dx}{\sin^2 ax \cos^2 ax} = -\frac{2}{a} \cot 2ax.$

365. $\displaystyle\int \frac{dx}{\sin^2 ax \cos^3 ax} = \frac{1}{a}\left[\frac{\sin ax}{2\cos^2 ax} - \frac{1}{\sin ax} + \frac{3}{2}\ln \tan\left(\frac{\pi}{4} + \frac{ax}{2}\right)\right].$

366. $\displaystyle\int \frac{dx}{\sin^3 ax \cos^2 ax} = \frac{1}{a}\left(\frac{1}{\cos ax} - \frac{\cos ax}{2\sin^2 ax} + \frac{3}{2}\ln \tan \frac{ax}{2}\right).$

367. $\displaystyle\int \frac{dx}{\sin ax \cos^n ax} = \frac{1}{a(n-1)\cos^{n-1} ax} + \int \frac{dx}{\sin ax \cos^{n-2} ax} \qquad (n \neq 1)$ (see No.361 and 363).

368. $\displaystyle\int \frac{dx}{\sin^n ax \cos ax} = -\frac{1}{a(n-1)\sin^{n-1} ax} + \int \frac{dx}{\sin^{n-2} ax \cos ax} \qquad (n \neq 1)$(see No. 360 and 362).

369. $$\int \frac{dx}{\sin^n ax \cos^m ax} = -\frac{1}{a(n-1)} \cdot \frac{1}{\sin^{n-1} ax \cos^{m-1} ax} + \frac{n+m-2}{n-1}\int \frac{dx}{\sin^{n-2} ax \cos^m ax}$$

(lowering the exponent n; $m > 0$, $n > 1$),

$$= \frac{1}{a(m-1)} \cdot \frac{1}{\sin^{n-1} ax \cos^{m-1} ax} + \frac{n+m-2}{n-1}\int \frac{dx}{\sin^n ax \cos^{m-2} ax}$$

(lowering the exponent m; $n > 0$, $m > 1$).

370. $\int \frac{\sin ax\,dx}{\cos^2 ax} = \frac{1}{a\cos ax} = \frac{1}{a}\sec ax.$

371. $\int \frac{\sin ax\,dx}{\cos^3 ax} = \frac{1}{2a\cos^2 ax} + C = \frac{1}{2a}\tan^2 ax + C_1.$

372. $\int \frac{\sin ax\,dx}{\cos^n ax} = \frac{1}{a(n-1)\cos^{n-1} ax}.$

373. $\int \frac{\sin^2 ax\,dx}{\cos ax} = -\frac{1}{a}\sin ax + \frac{1}{a}\ln\tan\left(\frac{\pi}{4} + \frac{ax}{2}\right).$

374. $\int \frac{\sin^2 ax\,dx}{\cos^3 ax} = \frac{1}{a}\left[\frac{\sin ax}{2\cos^2 ax} - \frac{1}{2}\ln\tan\left(\frac{\pi}{4} + \frac{ax}{2}\right)\right].$

375. $\int \frac{\sin^2 ax\,dx}{\cos^n ax} = \frac{\sin ax}{a(n-1)\cos^{n-1} ax} - \frac{1}{n-1}\int \frac{dx}{\cos^{n-2} ax} \qquad (n \neq 1)$ (see No. 325, 326, 328).

376. $\int \frac{\sin^3 ax\,dx}{\cos ax} = -\frac{1}{a}\left(\frac{\sin^2 ax}{2} + \ln\cos ax\right).$

377. $\int \frac{\sin^3 ax\,dx}{\cos^2 ax} = \frac{1}{a}\left(\cos ax + \frac{1}{\cos ax}\right).$

378. $\int \frac{\sin^3 ax\,dx}{\cos^n ax} = \frac{1}{a}\left[\frac{1}{(n-1)\cos^{n-1} ax} - \frac{1}{(n-3)\cos^{n-3} ax}\right] \qquad (n \neq 1,\ n \neq 3).$

379. $\int \frac{\sin^n ax}{\cos ax}dx = -\frac{\sin^{n-1} ax}{a(n-1)} + \int \frac{\sin^{n-2} ax\,dx}{\cos ax} \qquad (n \neq 1).$

380. $$\int \frac{\sin^n ax}{\cos^m ax}dx = = \frac{\sin^{n+1} ax}{a(m-1)\cos^{m-1} ax} - \frac{n-m+2}{m-1}\int \frac{\sin^n ax}{\cos^{m-2} ax}dx \qquad (m \neq 1),$$
$$= -\frac{\sin^{n-1} ax}{a(n-m)\cos^{m-1} ax} + \frac{n-1}{n-m}\int \frac{\sin^{n-2} ax\,dx}{\cos^m ax} \qquad (m \neq n),$$
$$= \frac{\sin^{n-1} ax}{a(m-1)\cos^{m-1} ax} - \frac{n-1}{m-1}\int \frac{\sin^{n-1} ax\,dx}{\cos^{m-2} ax} \qquad (m \neq 1).$$

381. $\int \frac{\cos ax\,dx}{\sin^2 ax} = -\frac{1}{a\sin ax} = -\frac{1}{a}\operatorname{cosec} ax.$

382. $\int \frac{\cos ax\,dx}{\sin^3 ax} = -\frac{1}{2a\sin^2 ax} + C = -\frac{\cot^2 ax}{2a} + C_1.$

383. $\int \frac{\cos ax\,dx}{\sin^n ax} = -\frac{1}{a(n-1)\sin^{n-1} ax}.$

384. $\int \frac{\cos^2 ax\,dx}{\sin ax} = \frac{1}{a}\left(\cos ax + \ln\tan\frac{ax}{2}\right).$

385. $\int \frac{\cos^2 ax\,dx}{\sin^3 ax} = -\frac{1}{2a}\left(\frac{\cos ax}{\sin^2 ax} - \ln\tan\frac{ax}{2}\right).$

386. $\displaystyle\int \frac{\cos^2 ax\,dx}{\sin^n ax} = -\frac{1}{(n-1)}\left(\frac{\cos ax}{a\sin^{n-1} ax} + \int \frac{dx}{\sin^{n-2} ax}\right) \qquad (n \neq 1)$ (see No. 289).

387. $\displaystyle\int \frac{\cos^3 ax\,dx}{\sin ax} = \frac{1}{a}\left(\frac{\cos^2 ax}{2} + \ln\sin ax\right).$

388. $\displaystyle\int \frac{\cos^3 ax\,dx}{\sin^2 ax} = -\frac{1}{a}\left(\sin ax + \frac{1}{\sin ax}\right).$

389. $\displaystyle\int \frac{\cos^3 ax\,dx}{\sin^n ax} = \frac{1}{a}\left[\frac{1}{(n-3)\sin^{n-3} ax} - \frac{1}{(n-1)\sin^{n-1} ax}\right] \qquad (n \neq 1, n \neq 3).$

390. $\displaystyle\int \frac{\cos^n ax}{\sin ax}\,dx = \frac{\cos^{n-1} ax}{a(n-1)} + \int \frac{\cos^{n-2} ax\,dx}{\sin ax} (n \neq 1).$

391.
$$\begin{aligned}
\int \frac{\cos^n ax\,dx}{\sin^m ax} &= -\frac{\cos^{n+1} ax}{a(m-1)\sin^{m-1} ax} - \frac{n-m+2}{m-1}\int \frac{\cos^n ax\,dx}{\sin^{m-2} ax} && (m \neq 1),\\
&= \frac{\cos^{n-1} ax}{a(n-m)\sin^{m-1} ax} + \frac{n-1}{m-1}\int \frac{\cos^{n-2} ax\,dx}{\sin^m ax} && (m \neq n),\\
&= -\frac{\cos^{n-1} ax}{a(m-1)\sin^{m-1} ax} - \frac{n-1}{m-1}\int \frac{\cos^{n-2} ax\,dx}{\sin^{m-2} ax} && (m \neq 1).
\end{aligned}$$

392. $\displaystyle\int \frac{dx}{\sin ax(1 \pm \cos ax)} = \pm\frac{1}{2a(1 \pm \cos ax)} + \frac{1}{2a}\ln\tan\frac{ax}{2}.$

393. $\displaystyle\int \frac{dx}{\cos ax(1 \pm \sin ax)} = \mp\frac{1}{2a(1 \pm \sin ax)} + \frac{1}{2a}\ln\tan\left(\frac{\pi}{4} + \frac{ax}{2}\right).$

394. $\displaystyle\int \frac{\sin ax\,dx}{\cos ax(1 \pm \cos ax)} = \frac{1}{a}\ln\frac{1 \pm \cos ax}{\cos ax}.$

395. $\displaystyle\int \frac{\cos ax\,dx}{\sin ax(1 \pm \sin ax)} = -\frac{1}{a}\ln\frac{1 \pm \sin ax}{\sin ax}.$

396. $\displaystyle\int \frac{\sin ax\,dx}{\cos ax(1 \pm \sin ax)} = \frac{1}{2a(1 \pm \sin ax)} \pm \frac{1}{2a}\ln\tan\left(\frac{\pi}{4} + \frac{ax}{2}\right).$

397. $\displaystyle\int \frac{\cos ax\,dx}{\sin ax(1 \pm \cos ax)} = -\frac{1}{2a(1 \pm \cos ax)} \pm \frac{1}{2a}\ln\tan\frac{ax}{2}.$

398. $\displaystyle\int \frac{\sin ax\,dx}{\sin ax \pm \cos ax} = \frac{x}{2} \mp \frac{1}{2a}\ln(\sin ax \pm \cos ax).$

399. $\displaystyle\int \frac{\cos ax\,dx}{\sin ax \pm \cos ax} = \pm\frac{x}{2} + \frac{1}{2a}\ln(\sin ax \pm \cos ax).$

400. $\displaystyle\int \frac{dx}{\sin ax \pm \cos ax} = \frac{1}{a\sqrt{2}}\ln\tan\left(\frac{ax}{2} \pm \frac{\pi}{8}\right).$

401. $\displaystyle\int \frac{dx}{1 + \cos ax \pm \sin ax} = \pm\frac{1}{a}\ln\left(1 \pm \tan\frac{ax}{2}\right).$

402. $\displaystyle\int \frac{dx}{b\sin ax + c\cos ax} = \frac{1}{a\sqrt{b^2 + c^2}}\ln\tan\frac{ax + \theta}{2} \qquad \text{with } \sin\theta = \frac{c}{\sqrt{b^2 + c^2}} \text{ and } \tan\theta = \frac{c}{b}.$

403. $\displaystyle\int \frac{\sin ax\,dx}{b + c\cos ax} = -\frac{1}{ac}\ln(b + c\cos ax).$

404. $\displaystyle\int \frac{\cos ax\,dx}{b + c\sin ax} = \frac{1}{ac}\ln(b + c\sin ax).$

405. $\displaystyle\int \frac{dx}{b + c\cos ax + f\sin ax} = \int \frac{d\left(x + \frac{\theta}{a}\right)}{b + \sqrt{c^2 + f^2}\sin(ax + \theta)}$

with $\sin\theta = \dfrac{c}{\sqrt{c^2 + f^2}}$ and $\tan\theta = \dfrac{c}{f}$ (see No. 306).

406. $\displaystyle\int \frac{dx}{b^2\cos^2 ax + c^2\sin^2 ax} = \frac{1}{abc}\arctan\left(\frac{c}{b}\tan ax\right).$

407. $\displaystyle\int \frac{dx}{b^2\cos^2 ax - c^2\sin^2 ax} = \frac{1}{2abc}\ln\frac{c\tan ax + b}{c\tan ax - b}.$

408. $\displaystyle\int \sin ax\cos bx\,dx = -\frac{\cos(a+b)x}{2(a+b)} - \frac{\cos(a-b)x}{2(a-b)}$ $(a^2 \neq b^2)$; for $a = b$ (see No. 354).

21.7.3.4 Integrals with Tangent Function

409. $\displaystyle\int \tan ax\,dx = -\frac{1}{a}\ln\cos ax.$

410. $\displaystyle\int \tan^2 ax\,dx = \frac{\tan ax}{a} - x.$

411. $\displaystyle\int \tan^3 ax\,dx = \frac{1}{2a}\tan^2 ax + \frac{1}{a}\ln\cos ax.$

412. $\displaystyle\int \tan^n ax\,dx = \frac{1}{a(n-1)}\tan^{n-1} ax - \int \tan^{n-2} ax\,dx.$

413. $\displaystyle\int x\tan ax\,dx = \frac{ax^3}{3} + \frac{a^3x^5}{15} + \frac{2a^5x^7}{105} + \frac{17a^7x^9}{2835} + \cdots + \frac{2^{2n}(2^{2n}-1)B_n a^{2n-1}x^{2n+1}}{(2n+1)!} + \cdots$

B_n denote the Bernoulli numbers (see 7.2.4.2, p. 465).

414. $\displaystyle\int \frac{\tan ax\,dx}{x} = ax + \frac{(ax)^3}{9} + \frac{2(ax)^5}{75} + \frac{17(ax)^7}{2205} + \cdots + \frac{2^{2n}(2^{2n}-1)B_n(ax)^{2n-1}}{(2n-1)(2n!)} + \cdots$

415. $\displaystyle\int \frac{\tan^n ax}{\cos^2 ax}\,dx = \frac{1}{a(n+1)}\tan^{n+1} ax \qquad (n \neq -1).$

416. $\displaystyle\int \frac{dx}{\tan ax \pm 1} = \pm\frac{x}{2} + \frac{1}{2a}\ln(\sin ax \pm \cos ax).$

417. $\displaystyle\int \frac{\tan ax\,dx}{\tan ax \pm 1} = \frac{x}{2} \mp \frac{1}{2a}\ln(\sin ax \pm \cos ax).$

21.7.3.5 Integrals with Cotangent Function

418. $\displaystyle\int \cot ax\,dx = \frac{1}{a}\ln\sin ax.$

419. $\int \cot^2 ax\,dx = -\frac{\cot ax}{a} - x.$

420. $\int \cot^3 ax\,dx = -\frac{1}{2a}\cot^2 ax - \frac{1}{a}\ln\sin ax.$

421. $\int \cot^n ax\,dx = -\frac{1}{a(n-1)}\cot^{n-1} ax - \int \cot^{n-2} ax\,dx \qquad (n \neq 1).$

422. $\int x\cot ax\,dx = \frac{x}{a} - \frac{ax^3}{9} - \frac{a^3x^5}{225} - \cdots - \frac{2^{2n}B_n a^{2n-1}x^{2n+1}}{(2n+1)!} - \cdots.$

B_n denote the Bernoulli numbers (see 7.2.4.2, p. 465).

423. $\int \frac{\cot ax\,dx}{x} = -\frac{1}{ax} - \frac{ax}{3} - \frac{(ax)^3}{135} - \frac{2(ax)^5}{4725} - \cdots - \frac{2^{2n}B_n(ax)^{2n-1}}{(2n-1)(2n)!} - \cdots.$

424. $\int \frac{\cot^n ax}{\sin^2 ax}dx = -\frac{1}{a(n+1)}\cot^{n+1} ax \qquad (n \neq -1).$

425. $\int \frac{dx}{1 \pm \cot ax} = \int \frac{\tan ax\,dx}{\tan ax \pm 1}$ (see No. 417).

21.7.4 Integrals of other Transcendental Functions

21.7.4.1 Integrals with Hyperbolic Functions

426. $\int \sinh ax\,dx = \frac{1}{a}\cosh ax.$

427. $\int \cosh ax\,dx = \frac{1}{a}\sinh ax.$

428. $\int \sinh^2 ax\,dx = \frac{1}{2a}\sinh ax\cosh ax - \frac{1}{2}x.$

429. $\int \cosh^2 ax\,dx = \frac{1}{2a}\sinh ax\cosh ax + \frac{1}{2}x.$

430. $\int \sinh^n ax\,dx$

$$= \frac{1}{an}\sinh^{n-1} ax\cosh ax - \frac{n-1}{n}\int \sinh^{n-2} ax\,dx \qquad (\text{for } n > 0),$$
$$= \frac{1}{a(n+1)}\sinh^{n+1} ax\cosh ax - \frac{n+2}{n+1}\int \sinh^{n+2} ax\,dx \qquad (\text{for } n < 0)\ (n \neq -1).$$

431. $\int \cosh^n ax\,dx$

$$= \frac{1}{an}\sinh ax\cosh^{n-1} ax + \frac{n-1}{n}\int \cosh^{n-2} ax\,dx \qquad (\text{for } n > 0),$$
$$= -\frac{1}{a(n+1)}\sinh ax\cosh^{n+1} ax + \frac{n+2}{n+1}\int \cosh^{n+2} ax\,dx \qquad (\text{for } n < 0)\ (n \neq -1).$$

432. $\int \frac{dx}{\sinh ax} = \frac{1}{a}\ln\tanh\frac{ax}{2}.$

433. $\int \frac{dx}{\cosh ax} = \frac{2}{a}\arctan e^{ax}.$

434. $\int x \sinh ax\, dx = \frac{1}{a} x \cosh ax - \frac{1}{a^2} \sinh ax.$

435. $\int x \cosh ax\, dx = \frac{1}{a} x \sinh ax - \frac{1}{a^2} \cosh ax.$

436. $\int \tanh ax\, dx = \frac{1}{a} \ln \cosh ax.$

437. $\int \coth ax\, dx = \frac{1}{a} \ln \sinh ax.$

438. $\int \tanh^2 ax\, dx = x - \frac{\tanh ax}{a}.$

439. $\int \coth^2 ax\, dx = x - \frac{\coth ax}{a}.$

440. $\int \sinh ax \sinh bx\, dx = \frac{1}{a^2 - b^2}(a \sinh bx \cosh ax - b \cosh bx \sinh ax) \qquad (a^2 \neq b^2).$

441. $\int \cosh ax \cosh bx\, dx = \frac{1}{a^2 - b^2}(a \sinh ax \cosh bx - b \sinh bx \cosh ax) \qquad (a^2 \neq b^2).$

442. $\int \cosh ax \sinh bx\, dx = \frac{1}{a^2 - b^2}(a \sinh bx \sinh ax - b \cosh bx \cosh ax) \qquad (a^2 \neq b^2).$

443. $\int \sinh ax \sin ax\, dx = \frac{1}{2a}(\cosh ax \sin ax - \sinh ax \cos ax).$

444. $\int \cosh ax \cos ax\, dx = \frac{1}{2a}(\sinh ax \cos ax + \cosh ax \sin ax).$

445. $\int \sinh ax \cos ax\, dx = \frac{1}{2a}(\cosh ax \cos ax + \sinh ax \sin ax).$

446. $\int \cosh ax \sin ax\, dx = \frac{1}{2a}(\sinh ax \sin ax - \cosh ax \cos ax).$

21.7.4.2 Integrals with Exponential Functions

447. $\int e^{ax}\, dx = \frac{1}{a} e^{ax}.$

448. $\int x e^{ax}\, dx = \frac{e^{ax}}{a^2}(ax - 1).$

449. $\int x^2 e^{ax}\, dx = e^{ax} \left(\frac{x^2}{a} - \frac{2x}{a^2} + \frac{2}{a^3} \right).$

450. $\int x^n e^{ax}\, dx = \frac{1}{a} x^n e^{ax} - \frac{n}{a} \int x^{n-1} e^{ax}\, dx.$

451. $\int \frac{e^{ax}}{x}\, dx = \ln x + \frac{ax}{1 \cdot 1!} + \frac{(ax)^2}{2 \cdot 2!} + \frac{(ax)^3}{3 \cdot 3!} + \cdots$

The definite integral $\int\limits_{-\infty}^{x} \frac{e^t}{t}\,dt$ is called the exponential function integral (see 8.2.5, **4.**, p. 514) and it is denoted by $\mathrm{Ei}(x)$. For $x > 0$ the integrand is divergent at $t = 0$; in this case we consider the principal value of the improper integral $\mathrm{Ei}(x)$ (see 8.2.5, **4.**, p. 514).

$$\int\limits_{-\infty}^{x} \frac{e^t}{t}\,dt = C + \ln|x| + \frac{x}{1\cdot 1!} + \frac{x^2}{2\cdot 2!} + \frac{x^3}{3\cdot 3!} + \cdots + \frac{x^n}{n\cdot n!} + \cdots .$$

C denotes the Euler constant (see 8.2.5, **2.**, p. 513).

452. $$\int \frac{e^{ax}}{x^n}\,dx = \frac{1}{n-1}\left(-\frac{e^{ax}}{x^{n-1}} + a\int \frac{e^{ax}}{x^{n-1}}dx\right) \qquad (n \neq 1).$$

453. $$\int \frac{dx}{1+e^{ax}} = \frac{1}{a}\ln\frac{e^{ax}}{1+e^{ax}}.$$

454. $$\int \frac{dx}{b+ce^{ax}} = \frac{x}{b} - \frac{1}{ab}\ln(b+ce^{ax}).$$

455. $$\int \frac{e^{ax}\,dx}{b+ce^{ax}} = \frac{1}{ac}\ln(b+ce^{ax}).$$

456. $$\int \frac{dx}{be^{ax}+ce^{-ax}} = \frac{1}{a\sqrt{bc}}\arctan\left(e^{ax}\sqrt{\frac{b}{c}}\right) \qquad (bc > 0),$$
$$= \frac{1}{2a\sqrt{-bc}}\ln\frac{c+e^{ax}\sqrt{-bc}}{c-e^{ax}\sqrt{-bc}} \qquad (bc < 0).$$

457. $$\int \frac{xe^{ax}\,dx}{(1+ax)^2} = \frac{e^{ax}}{a^2(1+ax)}.$$

458. $$\int e^{ax}\ln x\,dx = \frac{e^{ax}\ln x}{a} - \frac{1}{a}\int \frac{e^{ax}}{x}\,dx\,;$$ (see No. 451).

459. $$\int e^{ax}\sin bx\,dx = \frac{e^{ax}}{a^2+b^2}(a\sin bx - b\cos bx).$$

460. $$\int e^{ax}\cos bx\,dx = \frac{e^{ax}}{a^2+b^2}(a\cos bx + b\sin bx).$$

461. $$\int e^{ax}\sin^n x\,dx = \frac{e^{ax}\sin^{n-1}x}{a^2+n^2}(a\sin x - n\cos x) + \frac{n(n-1)}{a^2+n^2}\int e^{ax}\sin^{n-2}x\,dx\,;$$ (see No. 447 and 459).

462. $$\int e^{ax}\cos^n x\,dx = \frac{e^{ax}\cos^{n-1}x}{a^2+n^2}(a\cos x + n\sin x) + \frac{n(n-1)}{a^2+n^2}\int e^{ax}\cos^{n-2}x\,dx\,;$$ (see No. 447 and 460).

463. $$\int xe^{ax}\sin bx\,dx = \frac{xe^{ax}}{a^2+b^2}(a\sin bx - b\cos bx) - \frac{e^{ax}}{(a^2+b^2)^2}[(a^2-b^2)\sin bx - 2ab\cos bx].$$

464. $\int xe^{ax}\cos bx\,dx = \frac{xe^{ax}}{a^2+b^2}(a\cos bx + b\sin bx) - \frac{e^{ax}}{(a^2+b^2)^2}[(a^2-b^2)\cos bx + 2ab\sin bx].$

21.7.4.3 Integrals with Logarithmic Functions

465. $\int \ln x\,dx = x\ln x - x.$

466. $\int (\ln x)^2\,dx = x(\ln x)^2 - 2x\ln x + 2x.$

467. $\int (\ln x)^3\,dx = x(\ln x)^3 - 3x(\ln x)^2 + 6x\ln x - 6x.$

468. $\int (\ln x)^n\,dx = x(\ln x)^n - n\int (lnx)^{n-1}\,dx \qquad (n \neq -1).$

469. $\int \frac{dx}{\ln x} = \ln\ln x + \ln x + \frac{(\ln x)^2}{2\cdot 2!} + \frac{(\ln x)^3}{3\cdot 3!} + \cdots.$

The definite integral $\int\limits_0^x \frac{dt}{\ln t}$ is called the logarithm integral (see 8.2.5, p. 513) and it is denoted by $\mathrm{Li}(x)$.
For $x > 1$ the integrand is divergent at $t = 1$. In this case we consider the principal value of the improper integral $\mathrm{Li}(x)$ (see 8.2.5, p. 513).
The relation between the logarithm integral and the exponential function integral (see 8.2.5, p. 514) is: $\mathrm{Li}(x) = \mathrm{Ei}(\ln x)$.

470. $\int \frac{dx}{(\ln x)^n} = -\frac{x}{(n-1)(\ln x)^{n-1}} + \frac{1}{n-1}\int \frac{dx}{(\ln x)^{n-1}} \qquad (n \neq 1); \qquad$ (see No. 469).

471. $\int x^m \ln x\,dx = x^{m+1}\left[\frac{\ln x}{m+1} - \frac{1}{(m+1)^2}\right] \qquad (m \neq -1).$

472. $\int x^m(\ln x)^n\,dx = \frac{x^{m+1}(\ln x)^n}{m+1} - \frac{n}{m+1}\int x^m(\ln x)^{n-1}\,dx \quad (m \neq -1,\ n \neq -1; \quad$ (see No. 470).

473. $\int \frac{(\ln x)^n}{x}\,dx = \frac{(\ln x)^{n+1}}{n+1}.$

474. $\int \frac{\ln x}{x^m}\,dx = -\frac{\ln x}{(m-1)x^{m-1}} - \frac{1}{(m-1)^2x^{m-1}} \qquad (m \neq 1).$

475. $\int \frac{(\ln x)^n}{x^m}\,dx = -\frac{(\ln x)^n}{(m-1)x^{m-1}} + \frac{n}{m-1}\int \frac{(\ln x)^{n-1}}{x^m}dx \qquad (m \neq 1); \qquad$ (see No. 474).

476. $\int \frac{x^m\,dx}{\ln x} = \int \frac{e^{-y}}{y}\,dy \qquad$ with $y = -(m+1)\ln x; \qquad$ (see No. 451).

477. $\int \frac{x^m\,dx}{(\ln x)^n} = -\frac{x^{m+1}}{(n-1)(\ln x)^{n-1}} + \frac{m+1}{n-1}\int \frac{x^m\,dx}{(\ln x)^{n-1}} \qquad (n \neq 1).$

478. $\int \frac{dx}{x\ln x} = \ln\ln x.$

479. $\int \frac{dx}{x^n\ln x} = \ln\ln x - (n-1)\ln x + \frac{(n-1)^2(\ln x)^2}{2\cdot 2!} - \frac{(n-1)^3(\ln x)^3}{3\cdot 3!} + \cdots.$

480. $\displaystyle\int \frac{dx}{x(\ln x)^n} = \frac{-1}{(n-1)(\ln x)^{n-1}} \qquad (n \neq 1).$

481. $\displaystyle\int \frac{dx}{x^p(\ln x)^n} = \frac{-1}{x^{p-1}(n-1)(\ln x)^{n-1}} - \frac{p-1}{n-1}\int \frac{dx}{x^p(\ln x)^{n-1}} \qquad (n \neq 1).$

482. $\displaystyle\int \ln \sin x\, dx = x\ln x - x - \frac{x^3}{18} - \frac{x^5}{900} - \cdots - \frac{2^{2n-1}B_n x^{2n+1}}{n(2n+1)!} - \cdots .$

B_n denote the Bernoulli numbers (see 7.2.4.2, p. 465).

483. $\displaystyle\int \ln \cos x\, dx = -\frac{x^3}{6} - \frac{x^5}{60} - \frac{x^7}{315} - \cdots - \frac{2^{2n-1}(2^{2n}-1)B_n}{n(2n+1)!}x^{2n+1} - \cdots .$

484. $\displaystyle\int \ln \tan x\, dx = x\ln x - x + \frac{x^3}{9} + \frac{7x^5}{450} + \cdots + \frac{2^{2n}(2^{2n-1}-1)B_n}{n(2n+1)!}x^{2n+1} + \cdots .$

485. $\displaystyle\int \sin \ln x\, dx = \frac{x}{2}(\sin \ln x - \cos \ln x).$

486. $\displaystyle\int \cos \ln x\, dx = \frac{x}{2}(\sin \ln x + \cos \ln x).$

487. $\displaystyle\int e^{ax}\ln x\, dx = \frac{1}{a}e^{ax}\ln x - \frac{1}{a}\int \frac{e^{ax}}{x}dx\,;$ (see No. 451).

21.7.4.4 Integrals with Inverse Trigonometric Functions

488. $\displaystyle\int \arcsin\frac{x}{a}\, dx = x\arcsin\frac{x}{a} + \sqrt{a^2-x^2}.$

489. $\displaystyle\int x\arcsin\frac{x}{a}\, dx = \left(\frac{x^2}{2} - \frac{a^2}{4}\right)\arcsin\frac{x}{a} + \frac{x}{4}\sqrt{a^2-x^2}.$

490. $\displaystyle\int x^2\arcsin\frac{x}{a}\, dx = \frac{x^3}{3}\arcsin\frac{x}{a} + \frac{1}{9}(x^2+2a^2)\sqrt{a^2-x^2}.$

491. $\displaystyle\int \frac{\arcsin\frac{x}{a}\, dx}{x} = \frac{x}{a} + \frac{1}{2\cdot 3\cdot 3}\frac{x^3}{a^3} + \frac{1\cdot 3}{2\cdot 4\cdot 5\cdot 5}\frac{x^5}{a^5} + \frac{1\cdot 3\cdot 5}{2\cdot 4\cdot 6\cdot 7\cdot 7}\frac{x^7}{a^7} + \cdots .$

492. $\displaystyle\int \frac{\arcsin\frac{x}{a}\, dx}{x^2} = -\frac{1}{x}\arcsin\frac{x}{a} - \frac{1}{a}\ln\frac{a+\sqrt{a^2-x^2}}{x}.$

493. $\displaystyle\int \arccos\frac{x}{a}\, dx = x\arccos\frac{x}{a} - \sqrt{a^2-x^2}.$

494. $\displaystyle\int x\arccos\frac{x}{a}\, dx = \left(\frac{x^2}{2} - \frac{a^2}{4}\right)\arccos\frac{x}{a} - \frac{x}{4}\sqrt{a^2-x^2}.$

495. $\displaystyle\int x^2\arccos\frac{x}{a}\, dx = \frac{x^3}{3}\arccos\frac{x}{a} - \frac{1}{9}(x^2+2a^2)\sqrt{a^2-x^2}.$

496. $\displaystyle\int \frac{\arccos\frac{x}{a}\, dx}{x} = \frac{\pi}{2}\ln x - \frac{x}{a} - \frac{1}{2\cdot 3\cdot 3}\frac{x^3}{a^3} - \frac{1\cdot 3}{2\cdot 4\cdot 5\cdot 5}\frac{x^5}{a^5} - \frac{1\cdot 3\cdot 5}{2\cdot 4\cdot 6\cdot 7\cdot 7}\frac{x^7}{a^7} - \cdots .$

497. $\displaystyle\int \frac{\arccos\frac{x}{a}\, dx}{x^2} = -\frac{1}{x}\arccos\frac{x}{a} + \frac{1}{a}\ln\frac{a+\sqrt{a^2-x^2}}{x}.$

498. $\displaystyle\int \arctan\frac{x}{a}\,dx = x\arctan\frac{x}{a} - \frac{a}{2}\ln(a^2+x^2).$

499. $\displaystyle\int x\arctan\frac{x}{a}\,dx = \frac{1}{2}(x^2+a^2)\arctan\frac{x}{a} - \frac{ax}{2}.$

500. $\displaystyle\int x^2\arctan\frac{x}{a}\,dx = \frac{x^3}{3}\arctan\frac{x}{a} - \frac{ax^2}{6} + \frac{a^3}{6}\ln(a^2+x^2).$

501. $\displaystyle\int x^n\arctan\frac{x}{a}\,dx = \frac{x^{n+1}}{n+1}\arctan\frac{x}{a} - \frac{a}{n+1}\int\frac{x^{n+1}\,dx}{a^2+x^2} \qquad (n \neq -1).$

502. $\displaystyle\int\frac{\arctan\frac{x}{a}\,dx}{x} = \frac{x}{a} - \frac{x^3}{3^2a^3} + \frac{x^5}{5^2a^5} - \frac{x^7}{7^2a^7} + \cdots \qquad (|x| < |a|).$

503. $\displaystyle\int\frac{\arctan\frac{x}{a}\,dx}{x^2} = -\frac{1}{x}\arctan\frac{x}{a} - \frac{1}{2a}\ln\frac{a^2+x^2}{x^2}.$

504. $\displaystyle\int\frac{\arctan\frac{x}{a}\,dx}{x^n} = -\frac{1}{(n-1)x^{n-1}}\arctan\frac{x}{a} + \frac{a}{n-1}\int\frac{dx}{x^{n-1}(a^2+x^2)} \qquad (n \neq 1).$

505. $\displaystyle\int \operatorname{arccot}\frac{x}{a}\,dx = x\operatorname{arccot}\frac{x}{a} + \frac{a}{2}\ln(a^2+x^2).$

506. $\displaystyle\int x\operatorname{arccot}\frac{x}{a}\,dx = \frac{1}{2}(x^2+a^2)\operatorname{arccot}\frac{x}{a} + \frac{ax}{2}.$

507. $\displaystyle\int x^2\operatorname{arccot}\frac{x}{a}\,dx = \frac{x^3}{3}\operatorname{arccot}\frac{x}{a} + \frac{ax^2}{6} - \frac{a^3}{6}\ln(a^2+x^2).$

508. $\displaystyle\int x^n\operatorname{arccot}\frac{x}{a}\,dx = \frac{x^{n+1}}{n+1}\operatorname{arccot}\frac{x}{a} + \frac{a}{n+1}\int\frac{x^{n+1}dx}{a^2+x^2} \qquad (n \neq -1).$

509. $\displaystyle\int\frac{\operatorname{arccot}\frac{x}{a}\,dx}{x} = \frac{\pi}{2}\ln x - \frac{x}{a} + \frac{x^3}{3^2a^3} - \frac{x^5}{5^2a^5} - \frac{x^7}{7^2a^7} - \cdots.$

510. $\displaystyle\int\frac{\operatorname{arccot}\frac{x}{a}\,dx}{x^2} = -\frac{1}{x}\operatorname{arccot}\frac{x}{a} + \frac{1}{2a}\ln\frac{a^2+x^2}{x^2}.$

511. $\displaystyle\int\frac{\operatorname{arccot}\frac{x}{a}\,dx}{x^n} = -\frac{1}{(n-1)x^{n-1}}\operatorname{arccot}\frac{x}{a} - \frac{a}{n-1}\int\frac{dx}{x^{n-1}(a^2+x^2)} \qquad (n \neq 1).$

21.7.4.5 Integrals with Inverse Hyperbolic Functions

512. $\displaystyle\int \operatorname{Arsinh}\frac{x}{a}\,dx = x\operatorname{Arsinh}\frac{x}{a} - \sqrt{x^2+a^2}.$

513. $\displaystyle\int \operatorname{Arcosh}\frac{x}{a}\,dx = x\operatorname{Arcosh}\frac{x}{a} - \sqrt{x^2-a^2}.$

514. $\displaystyle\int \operatorname{Artanh}\frac{x}{a}\,dx = x\operatorname{Artanh}\frac{x}{a} + \frac{a}{2}\ln(a^2-x^2).$

515. $\displaystyle\int \operatorname{Arcoth}\frac{x}{a}\,dx = x\operatorname{Arcoth}\frac{x}{a} + \frac{a}{2}\ln(x^2-a^2).$

21.8 Definite Integrals

21.8.1 Definite Integrals of Trigonometric Functions

For natural numbers m, n:

1. $\int\limits_0^{2\pi} \sin nx\,dx = 0.$ (21.1) **2.** $\int\limits_0^{2\pi} \cos nx\,dx = 0.$ (21.2) **3.** $\int\limits_0^{2\pi} \sin nx \cos mx\,dx = 0.$ (21.3)

4. $\int\limits_0^{2\pi} \sin nx \sin mx\,dx = \begin{cases} 0 & \text{for } m \neq n, \\ \pi & \text{for } m = n. \end{cases}$ (21.4) **5.** $\int\limits_0^{2\pi} \cos nx \cos mx\,dx = \begin{cases} 0 & \text{for } m \neq n, \\ \pi & \text{for } m = n. \end{cases}$ (21.5)

6. $$\int\limits_0^{\frac{\pi}{2}} \sin^n x\,dx = \begin{cases} \frac{2}{3}\,\frac{4}{5}\,\frac{6}{7}\,\frac{8}{9}\cdots\frac{n-1}{n} & \text{for } n \text{ odd}, \\ \frac{\pi}{2}\,\frac{1}{2}\,\frac{3}{4}\,\frac{5}{6}\cdots\frac{n-1}{n} & \text{for } n \text{ even} \end{cases} \qquad (n \geq 2). \tag{21.6}$$

7a. $$\int\limits_0^{\pi/2} \sin^{2\alpha+1} x \cos^{2\beta+1} x\,dx = \frac{\Gamma(\alpha+1)\Gamma(\beta+1)}{2\Gamma(\alpha+\beta+2)} = \frac{1}{2}\mathrm{B}(\alpha+1, \beta+1). \tag{21.7a}$$

$\mathrm{B}(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$ denotes the *beta function* or the *Euler integral of the first kind*, $\Gamma(x)$ denotes the gamma function or the *Euler integral of the second kind* (see 8.2.5, **6.**, p. 514).
The formula (21.7a) is valid for arbitrary α and β; we use it, e.g., to determine the integrals

$$\int\limits_0^{\pi/2} \sqrt{\sin x}\,dx, \quad \int\limits_0^{\pi/2} \sqrt[3]{\sin x}\,dx, \quad \int\limits_0^{\pi/2} \frac{dx}{\sqrt[3]{\cos x}} \quad \text{etc.}$$

For positive integer α, β:

7b. $$\int\limits_0^{\pi/2} \sin^{2\alpha+1} x \cos^{2\beta+1} x\,dx = \frac{\alpha!\beta!}{2(\alpha+\beta+1)!}. \tag{21.7b}$$

8. $$\int\limits_0^{\infty} \frac{\sin ax}{x}\,dx = \begin{cases} \frac{\pi}{2} & \text{for } a > 0, \\ -\frac{\pi}{2} & \text{for } a < 0. \end{cases} \tag{21.8}$$

9. $$\int\limits_0^{\alpha} \frac{\cos ax\,dx}{x} = \infty \qquad (\alpha \text{ arbitrary}). \tag{21.9}$$

10. $$\int\limits_0^{\infty} \frac{\tan ax\,dx}{x} = \begin{cases} \frac{\pi}{2} & \text{for } a > 0, \\ -\frac{\pi}{2} & \text{for } a < 0. \end{cases} \tag{21.10}$$

11. $$\int\limits_0^{\infty} \frac{\cos ax - \cos bx}{x}\,dx = \ln\frac{b}{a}. \tag{21.11}$$

12. $$\int_0^\infty \frac{\sin x \cos ax}{x}\,dx = \begin{cases} \frac{\pi}{2} & \text{for } |a| < 1, \\ \frac{\pi}{4} & \text{for } |a| = 1, \\ 0 & \text{for } |a| > 1. \end{cases} \quad (21.12)$$

13. $$\int_0^\infty \frac{\sin x}{\sqrt{x}}\,dx = \int_0^\infty \frac{\cos x}{\sqrt{x}}\,dx = \sqrt{\frac{\pi}{2}}. \quad (21.13)$$

14. $$\int_0^\infty \frac{x \sin bx}{a^2 + x^2}\,dx = \pm\frac{\pi}{2}e^{-|ab|} \quad \text{(the sign is the same as the sign of } b\text{)}. \quad (21.14)$$

15. $$\int_0^\infty \frac{\cos ax}{1 + x^2}\,dx = \frac{\pi}{2}e^{-|a|}. \quad (21.15)$$

16. $$\int_0^\infty \frac{\sin^2 ax}{x^2}\,dx = \frac{\pi}{2}|a|. \quad (21.16)$$

17. $$\int_{-\infty}^{+\infty} \sin(x^2)\,dx = \int_{-\infty}^{+\infty} \cos(x^2)\,dx = \sqrt{\frac{\pi}{2}}. \quad (21.17)$$

18. $$\int_0^{\pi/2} \frac{\sin x\,dx}{\sqrt{1 - k^2 \sin^2 x}} = \frac{1}{2k}\ln\frac{1+k}{1-k} \quad \text{for } |k| < 1. \quad (21.18)$$

19. $$\int_0^{\pi/2} \frac{\cos x\,dx}{\sqrt{1 - k^2 \sin^2 x}} = \frac{1}{k}\arcsin k \quad \text{for } |k| < 1. \quad (21.19)$$

20. $$\int_0^{\pi/2} \frac{\sin^2 x\,dx}{\sqrt{1 - k^2 \sin^2 x}} = \frac{1}{k^2}(\mathrm{K} - \mathrm{E}) \quad \text{for } |k| < 1. \quad (21.20)$$

Here, and in the following, E and K mean complete elliptic integrals (see 8.1.4.3, **2.**, p. 490):
$\mathrm{E} = \mathrm{E}\left(k, \frac{\pi}{2}\right)$, $\mathrm{K} = F\left(k, \frac{\pi}{2}\right)$ (see also the table of elliptic integrals 21.9, p. 1103).

21. $$\int_0^{\pi/2} \frac{\cos^2 x\,dx}{\sqrt{1 - k^2 \sin^2 x}} = \frac{1}{k^2}[\mathrm{E} - (1 - k^2)\,\mathrm{K}]. \quad (21.21)$$

22. $$\int_0^{\pi} \frac{\cos ax\,dx}{1 - 2b\cos x + b^2} = \frac{\pi b^a}{1 - b^2} \quad \text{for integer } a \ge 0,\ |b| < 1. \quad (21.22)$$

21.8.2 Definite Integrals of Exponential Functions

(partially combined with algebraic, trigonometric, and logarithmic functions)

23. $$\int_0^\infty x^n e^{-ax}\,dx = \frac{\Gamma(n+1)}{a^{n+1}} \quad \text{for } a > 0,\ n > -1, \quad (21.23a)$$

$$= \frac{n!}{a^{n+1}} \quad \text{for } a > 0,\ n = 0, 1, 2, \ldots. \quad (21.23b)$$

$\Gamma(n)$ denotes the gamma function (see 8.2.5, **6.**, p. 514); see also the table of the gamma function 21.10, p. 1105).

24. $$\int_0^\infty x^n e^{-ax^2}\,dx = \frac{\Gamma\left(\frac{n+1}{2}\right)}{2a^{\left(\frac{n+1}{2}\right)}} \quad \text{for } a > 0,\ n > -1, \tag{21.24a}$$

$$= \frac{1\cdot 3\cdots(2k-1)\sqrt{\pi}}{2^{k+1}a^{k+1/2}} \quad \text{for } n = 2k\ (k = 1, 2, \ldots),\ a > 0, \tag{21.24b}$$

$$= \frac{k!}{2a^{k+1}} \quad \text{for } n = 2k+1\ (k = 0, 1, 2, \ldots),\ a > 0. \tag{21.24c}$$

25. $$\int_0^\infty e^{-a^2x^2}\,dx = \frac{\sqrt{\pi}}{2a} \quad \text{for } a > 0. \tag{21.25}$$

26. $$\int_0^\infty x^2 e^{-a^2x^2}\,dx = \frac{\sqrt{\pi}}{4a^3} \quad \text{for } a > 0. \tag{21.26}$$

27. $$\int_0^\infty e^{-a^2x^2}\cos bx\,dx = \frac{\sqrt{\pi}}{2a}\cdot e^{-b^2/4a^2} \quad \text{for } a > 0. \tag{21.27}$$

28. $$\int_0^\infty \frac{x\,dx}{e^x - 1} = \frac{\pi^2}{6}. \tag{21.28}$$

29. $$\int_0^\infty \frac{x\,dx}{e^x + 1} = \frac{\pi^2}{12}. \tag{21.29}$$

30. $$\int_0^\infty \frac{e^{-ax}\sin x}{x}dx = \operatorname{arccot} a = \arctan\frac{1}{a} \quad \text{for } a > 0. \tag{21.30}$$

31. $$\int_0^\infty e^{-x}\ln x\,dx = -C \approx -0{,}5772 \tag{21.31}$$

C denotes the Euler constant (see 8.2.5, **2.**, p. 513).

21.8.3 Definite Integrals of Logarithmic Functions

(combined with algebraic and trigonometric functions)

32. $$\int_0^1 \ln|\ln x|\ dx = -C = -0{,}5772 \quad \text{(reduced to Nr. 21.31).} \tag{21.32}$$

C is the Euler constant (see 8.2.5, **2.**, p. 513).

33. $$\int_0^1 \frac{\ln x}{x-1}\,dx = \frac{\pi^2}{6} \qquad \text{(reduced to Nr. 21.28).} \tag{21.33}$$

34. $$\int_0^1 \frac{\ln x}{x+1}\,dx = -\frac{\pi^2}{12} \quad \text{(reduced to Nr. 21.29).} \tag{21.34}$$

35. $$\int_0^1 \frac{\ln x}{x^2-1}\,dx = \frac{\pi^2}{8}\,. \qquad (21.35)$$

36. $$\int_0^1 \frac{\ln(1+x)}{x^2+1}\,dx = \frac{\pi}{8}\ln 2. \qquad (21.36)$$

37. $$\int_0^1 \left(\frac{1}{x}\right)^a dx = \Gamma(a+1) \quad \text{for } (-1 < a < \infty). \qquad (21.37)$$

$\Gamma(x)$ denotes the gamma function (see 8.2.5, **6.**, p. 514; see also the table of the gamma function 21.10, p. 1105).

38. $$\int_0^{\pi/2} \ln\sin x\,dx = \int_0^{\pi/2} \ln\cos x\,dx = -\frac{\pi}{2}\ln 2. \qquad (21.38)$$

39. $$\int_0^{\pi} x\ln\sin x\,dx = -\frac{\pi^2\ln 2}{2}\,. \qquad (21.39)$$

40. $$\int_0^{\pi/2} \sin x\ln\sin x\,dx = \ln 2 - 1. \qquad (21.40)$$

41. $$\int_0^{\pi} \ln(a \pm b\cos x)\,dx = \pi\ln\frac{a+\sqrt{a^2-b^2}}{2} \quad \text{for} a \ge b. \qquad (21.41)$$

42. $$\int_0^{\pi} \ln(a^2 - 2ab\cos x + b^2)\,dx = \begin{cases} 2\pi\ln a & \text{for} \quad (a \ge b > 0), \\ 2\pi\ln b & \text{for} \quad (b \ge a > 0). \end{cases} \qquad (21.42)$$

43. $$\int_0^{\pi/2} \ln\tan x\,dx = 0. \qquad (21.43)$$

44. $$\int_0^{\pi/4} \ln(1+\tan x)\,dx = \frac{\pi}{8}\ln 2. \qquad (21.44)$$

21.8.4 Definite Integrals of Algebraic Functions

45. $$\int_0^1 x^a(1-x)^\beta\,dx = 2\int_0^1 x^{2\alpha+1}(1-x^2)^\beta\,dx = \frac{\Gamma(\alpha+1)\Gamma(\beta+1)}{\Gamma(\alpha+\beta+2)}$$
$$= \mathrm{B}\,(\alpha+1,\ \beta+1), \quad \text{(reduced to Nr. 21.7a).} \qquad (21.45)$$

$\mathrm{B}(x,y) = \dfrac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$ denotes the beta function (see 21.8.1, p. 1098) or the Euler integral of the first kind, $\Gamma(x)$ denotes the gamma function (see 8.2.5, **6.**, p. 514) or the Euler integral of the second kind.

46. $$\int_0^{\infty} \frac{dx}{(1+x)x^a} = \frac{\pi}{\sin a\pi} \quad \text{for } a < 1. \qquad (21.46)$$

47. $$\int_0^\infty \frac{dx}{(1-x)x^a} = -\pi \cot a\pi \quad \text{for } a < 1. \tag{21.47}$$

48. $$\int_0^\infty \frac{x^{a-1}}{1+x^b}\,dx = \frac{\pi}{b \sin \dfrac{a\pi}{b}} \quad \text{for } 0 < a < b. \tag{21.48}$$

49. $$\int_0^1 \frac{dx}{\sqrt{1-x^a}} = \frac{\sqrt{\pi}\,\Gamma\left(\dfrac{1}{a}\right)}{a\,\Gamma\left(\dfrac{2+a}{2a}\right)}. \tag{21.49}$$

$\Gamma(x)$ denotes the gamma function (see 8.2.5, **6.**, p. 514; see also the table of the gamma function 21.10, p. 1105).

50. $$\int_0^1 \frac{dx}{1+2x\cos a + x^2} = \frac{a}{2\sin a} \quad \left(0 < a < \frac{\pi}{2}\right). \tag{21.50}$$

51. $$\int_0^\infty \frac{dx}{1+2x\cos a + x^2} = \frac{a}{\sin x} \quad \left(0 < a < \frac{\pi}{2}\right). \tag{21.51}$$

21.9 Elliptic Integrals

21.9.1 Elliptic Integral of the First Kind $F(\varphi, k)$, $k = \sin\alpha$

φ /°	α /°									
	0	10	20	30	40	50	60	70	80	90
0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
10	0.1745	0.1746	0.1746	0.1748	0.1749	0.1751	0.1752	0.1753	0.1754	0.1754
20	0.3491	0.3493	0.3499	0.3508	0.3520	0.3533	0.3545	0.3555	0.3561	0.3564
30	0.5236	0.5243	0.5263	0.5294	0.5334	0.5379	0.5422	0.5459	0.5484	0.5493
40	0.6981	0.6997	0.7043	0.7116	0.7213	0.7323	0.7436	0.7535	0.7604	0.7629
50	0.8727	0.8756	0.8842	0.8982	0.9173	0.9401	0.9647	0.9876	1.0044	1.0107
60	1.0472	1.0519	1.0660	1.0896	1.1226	1.1643	1.2126	1.2619	1.3014	1.3170
70	1.2217	1.2286	1.2495	1.2853	1.3372	1.4068	1.4944	1.5959	1.6918	1.7354
80	1.3963	1.4056	1.4344	1.4846	1.5597	1.6660	1.8125	2.0119	2.2653	2.4362
90	1.5708	1.5828	1.6200	1.6858	1.7868	1.9356	2.1565	2.5046	3.1534	∞

21.9.2 Elliptic Integral of the Second Kind $E(\varphi, k)$, $k = \sin\alpha$

φ /°	α /°									
	0	10	20	30	40	50	60	70	80	90
0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
10	0.1745	0.1745	0.1744	0.1743	0.1742	0.1740	0.1739	0.1738	0.1737	0.1736
20	0.3491	0.3489	0.3483	0.3473	0.3462	0.3450	0.3438	0.3429	0.3422	0.3420
30	0.5236	0.5229	0.5209	0.5179	0.5141	0.5100	0.5061	0.5029	0.5007	0.5000
40	0.6981	0.6966	0.6921	0.6851	0.6763	0.6667	0.6575	0.6497	0.6446	0.6428
50	0.8727	0.8698	0.8614	0.8483	0.8317	0.8134	0.7954	0.7801	0.7697	0.7660
60	1.0472	1.0426	1.0290	1.0076	0.9801	0.9493	0.9184	0.8914	0.8728	0.8660
70	1.2217	1.2149	1.1949	1.1632	1.1221	1.0750	1.0266	0.9830	0.9514	0.9397
80	1.3963	1.3870	1.3597	1.3161	1.2590	1.1926	1.1225	1.0565	1.0054	0.9848
90	1.5708	1.5589	1.5238	1.4675	1.3931	1.3055	1.2111	1.1184	1.0401	1.0000

21.9.3 Complete Elliptic Integral, $k = \sin\alpha$

α /°	K	E	α /°	K	E	α /°	K	E
0	1.5708	1.5708	**30**	1.6858	1.4675	**60**	2.1565	1.2111
1	1.5709	1.5707	31	1.6941	1.4608	61	2.1842	1.2015
2	1.5713	1.5703	32	1.7028	1.4539	62	2.2132	1.1920
3	1.5719	1.5697	33	1.7119	1.4469	63	2.2435	1.1826
4	1.5727	1.5689	34	1.7214	1.4397	64	2.2754	1.1732
5	1.5738	1.5678	**35**	1.7312	1.4323	**65**	2.3088	1.1638
6	1.5751	1.5665	36	1.7415	1.4248	66	2.3439	1.1545
7	1.5767	1.5649	37	1.7522	1.4171	67	2.3809	1.1453
8	1.5785	1.5632	38	1.7633	1.4092	68	2.4198	1.1362
9	1.5805	1.5611	39	1.7748	1.4013	69	2.4610	1.1272
10	1.5828	1.5589	**40**	1.7868	1.3931	**70**	2.5046	1.1184
11	1.5854	1.5564	41	1.7992	1.3849	71	2.5507	1.1096
12	1.5882	1.5537	42	1.8122	1.3765	72	2.5998	1.1011
13	1.5913	1.5507	43	1.8256	1.3680	73	2.6521	1.0927
14	1.5946	1.5476	44	1.8396	1.3594	74	2.7081	1.0844
15	1.5981	1.5442	**45**	1.8541	1.3506	**75**	2.7681	1.0764
16	1.6020	1.5405	46	1.8691	1.3418	76	2.8327	1.0686
17	1.6061	1.5367	47	1.8848	1.3329	77	2.9026	1.0611
18	1.6105	1.5326	48	1.9011	1.3238	78	2.9786	1.0538
19	1.6151	1.5283	49	1.9180	1.3147	79	3.0617	1.0468
20	1.6200	1.5238	**50**	1.9356	1.3055	**80**	3.1534	1.0401
21	1.6252	1.5191	51	1.9539	1.2963	81	3.2553	1.0338
22	1.6307	1.5141	52	1.9729	1.2870	82	3.3699	1.0278
23	1.6365	1.5090	53	1.9927	1.2776	83	3.5004	1.0223
24	1.6426	1.5037	54	2.0133	1.2681	84	3.6519	1.0172
25	1.6490	1.4981	**55**	2.0347	1.2587	**85**	3.8317	1.0127
26	1.6557	1.4924	56	2.0571	1.2492	86	4.0528	1.0080
27	1.6627	1.4864	57	2.0804	1.2397	87	4.3387	1.0053
28	1.6701	1.4803	58	2.1047	1.2301	88	4.7427	1.0026
29	1.6777	1.4740	59	2.1300	1.2206	89	5.4349	1.0008
						90	∞	1.0000

21.10 Gamma Function

x	$\Gamma(x)$	x	$\Gamma(x)$	x	$\Gamma(x)$	x	$\Gamma(x)$
1.00	1.00000	**1.25**	0.90640	**1.50**	0.88623	**1.75**	0.91906
01	0.99433	26	0.90440	51	0.88659	76	0.92137
02	0.98884	27	0.90250	52	0.88704	77	0.92376
03	0.98355	28	0.90072	53	0.88757	78	0.92623
04	0.97844	29	0.89904	54	0.88818	79	0.92877
1.05	0.97350	**1.30**	0.89747	**1.55**	0.88887	**1.80**	0.93138
06	0.96874	31	0.89600	56	0.88964	81	0.93408
07	0.96415	32	0.89464	57	0.89049	82	0.93685
08	0.95973	33	0.89338	58	0.89142	83	0.93969
09	0.95546	34	0.89222	59	0.89243	84	0.94261
1.10	0.95135	**1.35**	0.89115	**1.60**	0.89352	**1.85**	0.94561
11	0.94740	36	0.89018	61	0.89468	86	0.94869
12	0.94359	37	0.88931	62	0.89592	87	0.95184
13	0.93993	38	0.88854	63	0.89724	88	0.95507
14	0.93642	39	0.88785	64	0.89864	89	0.95838
1.15	0.93304	**1.40**	0.88726	**1.65**	0.90012	**1.90**	0.96177
16	0.92980	41	0.88676	66	0.90167	91	0.96523
17	0.92670	42	0.88636	67	0.90330	92	0.96877
18	0.92373	43	0.88604	68	0.90500	93	0.97240
19	0.92089	44	0.88581	69	0.90678	94	0.97610
1.20	0.91817	**1.45**	0.88566	**1.70**	0.90864	**1.95**	0.97988
21	0.91558	46	0.88560	71	0.91057	96	0.98374
22	0.91311	47	0.88563	72	0.91258	97	0.98768
23	0.91075	48	0.88575	73	0.91467	98	0.99171
24	0.90852	49	0.88592	74	0.91683	99	0.99581
1.25	0.90640	**1.50**	0.88623	**1.75**	0.91906	**2.00**	1.00000

The values of the gamma function for $x < 1$ $(x \neq 0, -1, -2, \ldots)$ and $x > 2$ can be calculated by the following formula:

$$\Gamma(x) = \frac{\Gamma(x+1)}{x}, \quad \Gamma(x) = (x-1)\,\Gamma(x-1).$$

■ **A:** $\Gamma(0.7) = \dfrac{\Gamma(1.7)}{0.7} = \dfrac{0.90864}{0.7} = 1.2981.$

■ **B:** $\Gamma(3.5) = 2.5 \cdot \Gamma(2.5) = 2.5 \cdot 1.5 \cdot \Gamma(1.5) = 2.5 \cdot 1.5 \cdot 0.88623 = 3.32336.$

21.11 Bessel Functions (Cylindrical Functions)

x	$J_0(x)$	$J_1(x)$	$Y_0(x)$	$Y_1(x)$	$I_0(x)$	$I_1(x)$	$K_0(x)$	$K_1(x)$
0.0	+1.0000	+0.0000	$-\infty$	$-\infty$	+1.000	0.0000	∞	∞
0.1	0.9975	0.0499	−1.5342	−6.4590	1.003	+0.0501	2.4271	9.8538
0.2	0.9900	0.0995	1.0181	3.3238	1.010	0.1005	1.7527	4.7760
0.3	0.9776	0.1483	0.8073	2.2931	1.023	0.1517	1.3725	3.0560
0.4	0.9604	0.1960	0.6060	1.7809	1.040	0.2040	1.1145	2.1844
0.5	+0.9385	+0.2423	−0.4445	−1.4715	1.063	0.2579	0.9244	1.6564
0.6	0.9120	0.2867	0.3085	1.2604	1.092	0.3137	0.7775	1.3028
0.7	0.8812	0.3290	0.1907	1.1032	1.126	0.3719	0.6605	1.0503
0.8	0.8463	0.3688	−0.0868	0.9781	1.167	0.4329	0.5653	0.8618
0.9	0.8075	0.4059	+0.0056	0.8731	1.213	0.4971	0.4867	0.7165
1.0	+0.7652	+0.4401	+0.0883	−0.7812	1.266	0.5652	0.4210	0.6019
1.1	0.7196	0.4709	0.1622	0.6981	1.326	0.6375	0.3656	0.5098
1.2	0.6711	0.4983	0.2281	0.6211	1.394	0.7147	0.3185	0.4346
1.3	0.6201	0.5220	0.2865	0.5485	1.469	0.7973	0.2782	0.3725
1.4	0.5669	0.5419	0.3379	0.4791	1.553	0.8861	0.2437	0.3208
1.5	+0.5118	+0.5579	+0.3824	−0.4123	1.647	0.9817	0.2138	0.2774
1.6	0.4554	0.5699	0.4204	0.3476	1.750	1.085	0.1880	0.2406
1.7	0.3980	0.5778	0.4520	0.2847	1.864	1.196	0.1655	0.2094
1.8	0.3400	0.5815	0.4774	0.2237	1.990	1.317	0.1459	0.1826
1.9	0.2818	0.5812	0.4968	0.1644	2.128	1.448	0.1288	0.1597
2.0	+0.2239	+0.5767	+0.5104	−0.1070	2.280	1.591	0.1139	0.1399
2.1	0.1666	0.5683	0.5183	−0.0517	2.446	1.745	0.1008	0.1227
2.2	0.1104	0.5560	0.5208	+0.0015	2.629	1.914	0.08927	0.1079
2.3	0.0555	0.5399	0.5181	0.0523	2.830	2.098	0.07914	0.09498
2.4	0.0025	0.5202	0.5104	0.1005	3.049	2.298	0.07022	0.08372
2.5	−0.0484	+0.4971	+0.4981	+0.1459	3.290	2.517	0.06235	0.07389
2.6	0.0968	0.4708	0.4813	0.1884	3.553	2.755	0.05540	0.06528
2.7	0.1424	0.4416	0.2605	0.2276	3.842	3.016	0.04926	0.05774
2.8	0.1850	0.4097	0.4359	0.2635	4.157	3.301	0.04382	0.05111
2.9	0.2243	0.3754	0.4079	0.2959	4.503	3.613	0.03901	0.04529
3.0	−0.2601	+0.3391	+0.3769	+0.3247	4.881	3.953	0.03474	0.04016
3.1	0.2921	0.3009	0.3431	0.3496	5.294	4.326	0.03095	0.03563
3.2	0.3202	0.2613	0.3070	0.3707	5.747	4.734	0.02759	0.03164
3.3	0.3443	0.2207	0.2691	0.3879	6.243	5.181	0.02461	0.02812
3.4	0.3643	0.1792	0.2296	0.4010	6.785	5.670	0.02196	0.02500
3.5	−0.3801	+0.1374	+0.1890	+0.4102	7.378	6.206	0.01960	0.02224
3.6	0.3918	0.0955	0.1477	0.4154	8.028	6.793	0.01750	0.01979
3.7	0.3992	0.0538	0.1061	0.4167	8.739	7.436	0.01563	0.01763
3.8	0.4026	+0.0128	0.0645	0.4141	9.517	8.140	0.01397	0.01571
3.9	0.4018	−0.0272	+0.0234	0.4078	10.37	8.913	0.01248	0.01400
4.0	−0.3971	−0.0660	−0.0169	+0.3979	11.30	9.759	0.01116	0.01248
4.1	0.3887	0.1033	0.0561	0.3846	12.32	10.69	0.009980	0.01114
4.2	0.3766	0.1386	0.0938	0.3680	13.44	11.71	0.008927	0.009938
4.3	0.3610	0.1719	0.1296	0.3484	14.67	12.82	0.007988	0.008872
4.4	0.3423	0.2028	0.1633	0.3260	16.01	14.05	0.007149	0.007923
4.5	−0.3205	−0.2311	−0.1947	+0.3010	17.48	15.39	0.006400	0.007078
4.6	0.2961	0.2566	0.2235	0.2737	19.09	16.86	0.005730	0.006325
4.7	0.2693	0.2791	0.2494	0.2445	20.86	18.48	0.005132	0.005654
4.8	0.2404	0.2985	0.2723	0.2136	22.79	20.25	0.004597	0.005055
4.9	0.2097	0.3147	0.2921	0.1812	24.91	22.20	0.004119	0.004521

x	$J_0(x)$	$J_1(x)$	$Y_0(x)$	$Y_1(x)$	$I_0(x)$	$I_1(x)$	$K_0(x)$	$K_1(x)$
							0.00	0.00
5.0	−0.1776	−0.3276	−0.3085	+0.1479	27.24	24.34	3691	4045
5.1	0.1443	0.3371	0.3216	0.1137	29.79	26.68	3308	3619
5.2	0.1103	0.3432	0.3313	0.0792	32.58	29.25	2966	3239
5.3	0.0758	0.3460	0.3374	0.0445	35.65	32.08	2659	2900
5.4	0.0412	0.3453	0.3402	+0.0101	39.01	35.18	2385	2597
5.5	−0.0068	−0.3414	−0.3395	−0.0238	42.69	38.59	2139	2326
5.6	+0.0270	0.3343	0.3354	0.0568	46.74	42.33	1918	2083
5.7	0.0599	0.3241	0.3282	0.0887	51.17	46.44	1721	1866
5.8	0.0917	0.3110	0.3177	0.1192	56.04	50.95	1544	1673
5.9	0.1220	0.2951	0.3044	0.1481	61.38	55.90	1386	1499
6.0	+0.1506	−0.2767	−0.2882	−0.1750	67.23	61.34	1244	1344
6.1	0.1773	0.2559	0.2694	0.1998	73.66	67.32	1117	1205
6.2	0.2017	0.2329	0.2483	0.2223	80.72	73.89	1003	1081
6.3	0.2238	0.2081	0.2251	0.2422	88.46	81.10	09001	09691
6.4	0.2433	0.1816	0.1999	0.2596	96.96	89.03	08083	08693
6.5	+0.2601	−0.1538	−0.1732	−0.2741	106.3	97.74	07259	07799
6.6	0.2740	0.1250	0.1452	0.2857	116.5	107.3	06520	06998
6.7	0.2851	0.0953	0.1162	0.2945	127.8	117.8	05857	06280
6.8	0.2931	0.0652	0.0864	0.3002	140.1	129.4	05262	05636
6.9	0.2981	0.0349	0.0563	0.3029	153.7	142.1	04728	05059
7.0	+0.3001	−0.0047	−0.0259	−0.3027	168.6	156.0	04248	04542
7.1	0.2991	+0.0252	+0.0042	0.2995	185.0	171.4	03817	04078
7.2	0.2951	0.0543	0.0339	0.2934	202.9	188.3	03431	03662
7.3	0.2882	0.0826	0.0628	0.2846	222.7	206.8	03084	03288
7.4	0.2786	0.1096	0.0907	0.2731	244.3	227.2	02772	02953
7.5	+0.2663	+0.1352	+0.1173	−0.2591	268.2	249.6	02492	02653
7.6	0.2516	0.1592	0.1424	0.2428	294.3	274.2	02240	02383
7.7	0.2346	0.1813	0.1658	0.2243	323.1	301.3	02014	02141
7.8	0.2154	0.2014	0.1872	0.2039	354.7	331.1	01811	01924
7.9	0.1944	0.2192	0.2065	0.1817	389.4	363.9	01629	01729
8.0	+0.1717	+0.2346	+0.2235	−0.1581	427.6	399.9	01465	01554
8.1	0.1475	0.2476	0.2381	0.1331	469.5	439.5	01317	01396
8.2	0.1222	0.2580	0.2501	0.1072	515.6	483.0	01185	01255
8.3	0.0960	0.2657	0.2595	0.0806	566.3	531.0	01066	01128
8.4	0.0692	0.2708	0.2662	0.0535	621.9	583.7	009588	01014
8.5	+0.0419	+0.2731	+0.2702	−0.0262	683.2	641.6	008626	009120
8.6	+0.0146	0.2728	0.2715	+0.0011	750.5	705.4	007761	008200
8.7	−0.0125	0.2697	0.2700	0.0280	824.4	775.5	006983	007374
8.8	0.0392	0.2641	0.2659	0.0544	905.8	852.7	006283	006631
8.9	0.0653	0.2559	0.2592	0.0799	995.2	937.5	005654	005964
9.0	−0.0903	+0.2453	+0.2499	+0.1043	1094	1031	005088	005364
9.1	0.1142	0.2324	0.2383	0.1275	1202	1134	004579	004825
9.2	0.1367	0.2174	0.2245	0.1491	1321	1247	004121	004340
9.3	0.1577	0.2004	0.2086	0.1691	1451	1371	003710	003904
9.4	0.1768	0.1816	0.1907	0.1871	1595	1508	003339	003512
9.5	−0.1939	+0.1613	+0.1712	+0.2032	1753	1658	003036	003160
9.6	0.2090	0.1395	0.1502	0.2171	1927	1824	002706	002843
9.7	0.2218	0.1166	0.1279	0.2287	2119	2006	002436	002559
9.8	0.2323	0.0928	0.1045	0.2379	2329	2207	002193	002302
9.9	0.2403	0.0684	0.0804	0.2447	2561	2428	001975	002072
10.0	−0.2459	+0.0435	+0.0557	+0.2490	2816	2671	001778	001865

21.12 Legendre Polynomials of the First Kind

$P_0(x) = 1;$ $\qquad P_1(x) = x;$

$P_2(x) = \frac{1}{2}(3x^2 - 1);$ $\qquad P_3(x) = \frac{1}{2}(5x^3 - 3x);$

$P_4(x) = \frac{1}{8}(35x^4 - 30x^2 + 3);$ $\qquad P_5(x) = \frac{1}{8}(63x^5 - 70x^3 + 15x);$

$P_6(x) = \frac{1}{16}(231x^6 - 315x^4 + 105x^2 - 5);$ $\qquad P_7(x) = \frac{1}{16}(429x^7 - 693x^5 + 315x^3 - 35x).$

$x = P_1(x)$	$P_2(x)$	$P_3(x)$	$P_4(x)$	$P_5(x)$	$P_6(x)$	$P_7(x)$
0.00	−0.3000	0.0000	0.3750	0.0000	−0.3125	0.0000
0.05	−0.4962	−0.0747	0.3657	0.0927	−0.2962	−0.1069
0.10	−0.4850	−0.1475	0.3379	0.1788	−0.2488	−0.1995
0.15	−0.4662	−0.2166	0.2928	0.2523	−0.1746	−0.2649
0.20	−0.4400	−0.2800	0.2320	0.3075	−0.0806	−0.2935
0.25	−0.4062	−0.3359	0.1577	0.3397	+0.0243	−0.2799
0.30	−0.3650	−0.3825	+0.0729	0.3454	0.1292	−0.2241
0.35	−0.3162	−0.4178	−0.0187	0.3225	0.2225	−0.1318
0.40	−0.2600	−0.4400	−0.1130	0.2706	0.2926	−0.0146
0.45	−0.1962	−0.4472	−0.2050	0.1917	0.3290	+0.1106
0.50	−0.1250	−0.4375	−0.2891	+0.0898	0.3232	0.2231
0.55	−0.0462	−0.4091	−0.3590	−0.0282	0.2708	0.3007
0.60	+0.0400	−0.3600	−0.4080	−0.1526	0.1721	0.3226
0.65	0.1338	−0.2884	−0.4284	−0.2705	+0.0347	0.2737
0.70	0.2350	−0.1925	−0.4121	−0.3652	−0.1253	+0.1502
0.75	0.3438	−0.0703	−0.3501	−0.4164	−0.2808	−0.0342
0.80	0.4600	+0.0800	−0.2330	−0.3995	−0.3918	−0.2397
0.85	0.5838	0.2603	−0.0506	−0.2857	−0.4030	−0.3913
0.90	0.7150	0.4725	+0.2079	−0.0411	−0.2412	−0.3678
0.95	0.8538	0.7184	0.5541	+0.3727	+0.1875	+0.0112
1.00	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

21.13 Laplace Transformation

(see 15.2.1.1, p. 770)

$$F(p) = \int_0^\infty e^{-pt} f(t)\, dt, \quad f(t) = 0 \text{ for } t < 0.$$

C is the Euler constant: $C = 0.577216$ (see 8.2.5, **2.**, p. 513).

No.	$F(p)$	$f(t)$
1	0	0
2	$\frac{1}{p}$	1
3	$\frac{1}{p^n}$	$\frac{t^{n-1}}{(n-1)!}$
4	$\frac{1}{(p-\alpha)^n}$	$\frac{t^{n-1}}{(n-1)!}e^{\alpha t}$
5	$\frac{1}{(p-\alpha)(p-\beta)}$	$\frac{e^{\beta t}-e^{\alpha t}}{\beta-\alpha}$
6	$\frac{p}{(p-\alpha)(p-\beta)}$	$\frac{\beta e^{\beta t}-\alpha e^{\alpha t}}{\beta-\alpha}$
7	$\frac{1}{p^2+2\alpha p+\beta^2}$	$\frac{e^{-\alpha t}}{\sqrt{\beta^2-\alpha^2}}\sin\sqrt{\beta^2-\alpha^2}\,t$
8	$\frac{\alpha}{p^2+\alpha^2}$	$\sin\alpha t$
9	$\frac{\alpha\cos\beta+p\sin\beta}{p^2+\alpha^2}$	$\sin(\alpha t+\beta)$
10	$\frac{p}{p^2+2\alpha p+\beta^2}$	$\left(\cos\sqrt{\beta^2-\alpha^2}\,t-\frac{\alpha}{\sqrt{\beta^2-\alpha^2}}\sin\sqrt{\beta^2-\alpha^2}\,t\right)e^{-\alpha t}$
11	$\frac{p}{p^2+\alpha^2}$	$\cos\alpha t$
12	$\frac{p\cos\beta-\alpha\sin\beta}{p^2+\alpha^2}$	$\cos(\alpha t+\beta)$
13	$\frac{\alpha}{p^2-\alpha^2}$	$\sinh\alpha t$
14	$\frac{p}{p^2-\alpha^2}$	$\cosh\alpha t$
15	$\frac{1}{(p-\alpha)(p-\beta)(p-\gamma)}$	$-\frac{(\beta-\gamma)e^{\alpha t}+(\gamma-\alpha)e^{\beta t}+(\alpha-\beta)e^{\gamma t}}{(\alpha-\beta)(\beta-\gamma)(\gamma-\alpha)}$

No.	$F(p)$	$f(t)$
16	$\frac{1}{(p-\alpha)(p-\beta)^2}$	$\frac{e^{\alpha t} - [1+(\alpha-\beta)t]\,e^{\beta t}}{(\alpha-\beta)^2}$
17	$\frac{p}{(p-\alpha)(p-\beta)^2}$	$\frac{\alpha\, e^{\alpha t} - [\alpha+\beta(\alpha-\beta)t]e^{\beta t}}{(\alpha-\beta)^2}$
18	$\frac{p^2}{(p-\alpha)(p-\beta)^2}$	$\frac{\alpha^2 e^{\alpha t} - [2\alpha-\beta+\beta(\alpha-\beta)t]\,\beta e^{\beta t}}{(\alpha-\beta)^2}$
19	$\frac{1}{(p^2+\alpha^2)(p^2+\beta^2)}$	$\frac{\alpha\sin\beta t - \beta\sin\alpha t}{\alpha\beta(\alpha^2-\beta^2)}$
20	$\frac{p}{(p^2+\alpha^2)(p^2+\beta^2)}$	$\frac{\cos\beta t - \cos\alpha t}{(\alpha^2-\beta^2)}$
21	$\frac{p^2+2\alpha^2}{p(p^2+4\alpha^2)}$	$\cos^2\alpha t$
22	$\frac{2\alpha^2}{p(p^2+4\alpha^2)}$	$\sin^2\alpha t$
23	$\frac{p^2-2\alpha^2}{p(p^2-4\alpha^2)}$	$\cosh^2\alpha t$
24	$\frac{2\alpha^2}{p(p^2-4\alpha^2)}$	$\sinh^2\alpha t$
25	$\frac{2\alpha^2 p}{p^4+4\alpha^4}$	$\sin\alpha t\cdot\sinh\alpha t$
26	$\frac{\alpha(p^2+2\alpha^2)}{p^4+4\alpha^4}$	$\sin\alpha t\cdot\cosh\alpha t$
27	$\frac{\alpha(p^2-2\alpha^2)}{p^4+4\alpha^4}$	$\cos\alpha t\cdot\sinh\alpha t$
28	$\frac{p^3}{p^4+4\alpha^4}$	$\cos\alpha t\cdot\cosh\alpha t$
29	$\frac{\alpha p}{(p^2+\alpha^2)^2}$	$\frac{t}{2}\sin\alpha t$
30	$\frac{\alpha p}{(p^2-\alpha^2)^2}$	$\frac{t}{2}\sinh\alpha t$
31	$\frac{\alpha\beta}{(p^2-\alpha^2)(p^2-\beta^2)}$	$\frac{\beta\sinh\alpha t - \alpha\sinh\beta t}{\alpha^2-\beta^2}$

No.	$F(p)$	$f(t)$
32	$\dfrac{p}{(p^2-\alpha^2)(p^2-\beta^2)}$	$\dfrac{\cosh\alpha t-\cosh\beta t}{\alpha^2-\beta^2}$
33	$\dfrac{1}{\sqrt{p}}$	$\dfrac{1}{\sqrt{\pi t}}$
34	$\dfrac{1}{p\sqrt{p}}$	$2\sqrt{\dfrac{t}{\pi}}$
35	$\dfrac{1}{p^n\sqrt{p}}$	$\dfrac{n!}{(2n)!}\dfrac{4^n}{\sqrt{\pi}}t^{n-\frac{1}{2}}\quad (n>0\,,\ \text{integer})$
36	$\dfrac{1}{\sqrt{p+\alpha}}$	$\dfrac{1}{\sqrt{\pi t}}\mathrm{e}^{-\alpha t}$
37	$\sqrt{p+\alpha}-\sqrt{p+\beta}$	$\dfrac{1}{2t\sqrt{\pi t}}\left(\mathrm{e}^{-\beta t}-\mathrm{e}^{-\alpha t}\right)$
38	$\sqrt{\sqrt{p^2+\alpha^2}-p}$	$\dfrac{\sin\alpha t}{t\sqrt{2\pi t}}$
39	$\sqrt{\dfrac{\sqrt{p^2+\alpha^2}-p}{p^2+\alpha^2}}$	$\sqrt{\dfrac{2}{\pi t}}\sin\alpha t$
40	$\sqrt{\dfrac{\sqrt{p^2+\alpha^2}+p}{p^2+\alpha^2}}$	$\sqrt{\dfrac{2}{\pi t}}\cos\alpha t$
41	$\sqrt{\dfrac{\sqrt{p^2-\alpha^2}-p}{p^2-\alpha^2}}$	$\sqrt{\dfrac{2}{\pi t}}\sinh\alpha t$
42	$\sqrt{\dfrac{\sqrt{p^2-\alpha^2}+p}{p^2-\alpha^2}}$	$\sqrt{\dfrac{2}{\pi t}}\cosh\alpha t$
43	$\dfrac{1}{p\sqrt{p+\alpha}}$	$\dfrac{2}{\sqrt{\alpha\pi}}\cdot\displaystyle\int_0^{\sqrt{\alpha t}}\mathrm{e}^{-\tau^2}\mathrm{d}\tau$
44	$\dfrac{1}{(p+\alpha)\sqrt{p+\beta}}$	$\dfrac{2\mathrm{e}^{-\alpha t}}{\sqrt{\pi(\beta-\alpha)}}\cdot\displaystyle\int_0^{\sqrt{(\beta-\alpha)t}}\mathrm{e}^{-\tau^2}\mathrm{d}\tau$
45	$\dfrac{\sqrt{p+\alpha}}{p}$	$\dfrac{\mathrm{e}^{-\alpha t}}{\sqrt{\pi t}}+2\sqrt{\dfrac{\alpha}{\pi}}\cdot\displaystyle\int_0^{\sqrt{\alpha t}}\mathrm{e}^{-\tau^2}\mathrm{d}\tau$
46	$\dfrac{1}{\sqrt{p^2+\alpha^2}}$	$J_0(\alpha t)$ (Bessel function of order 0, p. 562)
47	$\dfrac{1}{\sqrt{p^2-\alpha^2}}$	$I_0(\alpha t)$ (modified Bessel function of order 0, p. 562)

No.	$F(p)$	$f(t)$
48	$\dfrac{1}{\sqrt{(p+\alpha)(p+\beta)}}$	$e^{-\frac{\alpha+\beta}{2}t} \cdot I_0\left(\dfrac{\alpha-\beta}{2}\,t\right)$
49	$\dfrac{1}{\sqrt{p^2+2\alpha p+\beta^2}}$	$e^{-\alpha t} \cdot J_0\left(\sqrt{\alpha^2-\beta^2}\,t\right)$
50	$\dfrac{e^{1/p}}{p\sqrt{p}}$	$\dfrac{\sinh 2\sqrt{t}}{\sqrt{\pi}}$
51	$\arctan\dfrac{\alpha}{p}$	$\dfrac{\sin\alpha t}{t}$
52	$\arctan\dfrac{2\alpha p}{p^2-\alpha^2+\beta^2}$	$\dfrac{2}{t}\sin\alpha t\cdot\cos\beta t$
53	$\arctan\dfrac{p^2-\alpha p+\beta}{\alpha\beta}$	$\dfrac{e^{\alpha t}-1}{t}\sin\beta t$
54	$\dfrac{\ln p}{p}$	$-C-\ln t$
55	$\dfrac{\ln p}{p^{n+1}}$	$\dfrac{t^n}{n!}[\psi(n)-\ln t]\,,\quad \psi(n)=1+\dfrac{1}{2}+\cdots+\dfrac{1}{n}-C$
56	$\dfrac{(\ln p)^2}{p}$	$(\ln t+C)^2-\dfrac{\pi^2}{6}$
57	$\ln\dfrac{p-\alpha}{p-\beta}$	$\dfrac{1}{t}\left(e^{\beta t}-e^{\alpha t}\right)$
58	$\ln\dfrac{p+\alpha}{p-\alpha}=2\operatorname{artanh}\dfrac{\alpha}{p}$	$\dfrac{2}{t}\sinh\alpha t$
59	$\ln\dfrac{p^2+\alpha^2}{p^2+\beta^2}$	$2\cdot\dfrac{\cos\beta t-\cos\alpha t}{t}$
60	$\ln\dfrac{p^2-\alpha^2}{p^2-\beta^2}$	$2\cdot\dfrac{\cosh\beta t-\cosh\alpha t}{t}$
61	$e^{-\alpha\sqrt{p}}\,,\quad \operatorname{Re}\alpha>0$	$\dfrac{\alpha}{2\sqrt{\pi}}\,\dfrac{e^{-\alpha^2/4t}}{t\sqrt{t}}$
62	$\dfrac{1}{\sqrt{p}}e^{-\alpha\sqrt{p}}\,,\quad \operatorname{Re}\alpha\geq 0$	$\dfrac{e^{-\alpha^2/4t}}{\sqrt{\pi t}}$
63	$\dfrac{\left(\sqrt{p^2+\alpha^2}-p\right)^\nu}{\sqrt{p^2+\alpha^2}}\,,\quad \operatorname{Re}\nu>-1$	$\alpha^\nu J_\nu(\alpha t)$ (see Bessel function, p. 562)

No.	$F(p)$	$f(t)$
64	$\dfrac{\left(p-\sqrt{p^2-\alpha^2}\right)^\nu}{\sqrt{p^2-\alpha^2}}, \quad \operatorname{Re}\nu > -1$	$\alpha^\nu I_\nu(\alpha t)$ (see Bessel function, p. 563)
65	$\dfrac{1}{p}\mathrm{e}^{-\beta p} \quad (\beta > 0\,,\ \text{reell})$	$\begin{cases} 0 & \text{for } t < \beta \\ 1 & \text{for } t > \beta \end{cases}$
66	$\dfrac{\mathrm{e}^{-\beta\sqrt{p^2+\alpha^2}}}{\sqrt{p^2+\alpha^2}}$	$\begin{cases} 0 & \text{for } t < \beta \\ J_0\left(\alpha\sqrt{t^2-\beta^2}\right) & \text{for } t > \beta \end{cases}$
67	$\dfrac{\mathrm{e}^{-\beta\sqrt{p^2-\alpha^2}}}{\sqrt{p^2-\alpha^2}}$	$\begin{cases} 0 & \text{for } t < \beta \\ I_0\left(\alpha\sqrt{t^2-\beta^2}\right) & \text{for } t > \beta \end{cases}$
68	$\dfrac{\mathrm{e}^{-\beta\sqrt{(p+\alpha)(p+\beta)}}}{\sqrt{(p+\alpha)(p+\beta)}}$	$\begin{cases} 0 & \text{for } t < \beta \\ \mathrm{e}^{-(\alpha+\beta)\frac{t}{2}} I_0\left(\dfrac{\alpha-\beta}{2}\sqrt{t^2-\beta^2}\right) & \text{for } t > \beta \end{cases}$
69	$\dfrac{\mathrm{e}^{-\beta\sqrt{p^2+\alpha^2}}}{p^2+\alpha^2}\left(\beta+\dfrac{1}{\sqrt{p^2+\alpha^2}}\right)$	$\begin{cases} 0 & \text{for } t < \beta \\ \dfrac{\sqrt{t^2-\beta^2}}{\alpha} J_1\left(\alpha\sqrt{t^2-\beta^2}\right) & \text{for } t > \beta \end{cases}$
70	$\dfrac{\mathrm{e}^{-\beta\sqrt{p^2-\alpha^2}}}{p^2-\alpha^2}\left(\beta+\dfrac{1}{\sqrt{p^2-\alpha^2}}\right)$	$\begin{cases} 0 & \text{for } t < \beta \\ \dfrac{\sqrt{t^2-\beta^2}}{\alpha} I_1\left(\alpha\sqrt{t^2-\beta^2}\right) & \text{for } t > \beta \end{cases}$
71	$\mathrm{e}^{-\beta p}-\mathrm{e}^{-\beta\sqrt{p^2+\alpha^2}}$	$\begin{cases} 0 & \text{for } t < \beta \\ \dfrac{\beta\alpha}{\sqrt{t^2-\beta^2}} J_1\left(\alpha\sqrt{t^2-\beta^2}\right) & \text{for } t > \beta \end{cases}$
72	$\mathrm{e}^{-\beta\sqrt{p^2-\alpha^2}}-\mathrm{e}^{-\beta p}$	$\begin{cases} 0 & \text{for } t < \beta \\ \dfrac{\beta\alpha}{\sqrt{t^2-\beta^2}} I_1\left(\alpha\sqrt{t^2-\beta^2}\right) & \text{for } t > \beta \end{cases}$
73	$\dfrac{1-\mathrm{e}^{-\alpha p}}{p}$	$\begin{cases} 0 \text{ for } t > \alpha \\ 1 \text{ for } 0 < t < \alpha \end{cases}$
74	$\dfrac{\mathrm{e}^{-\alpha p}-\mathrm{e}^{-\beta p}}{p}$	$\begin{cases} 0 \text{ for } 0 < t < \alpha \\ 1 \text{ for } \alpha < t < \beta \\ 0 \text{ for } t > \beta \end{cases}$

21.14 Fourier Transformation

The symbols in the table are defined in the following way:

C: Euler constant ($C = 0.577215\ldots$)

$$\Gamma(z) = \int_0^\infty e^{-t} t^{z-1}\, dt, \quad \operatorname{Re} z > 0$$ (Gamma function, see 8.2.5, **6.**, p. 514),

$$J_\nu(z) = \sum_{n=0}^{\infty} \frac{(-1)^n (\frac{1}{2}z)^{\nu+2n}}{n!\,\Gamma(\nu+n+1)}$$ (Bessel functions, see 9.1.2.6, **2.**, p. 562),

$$K_\nu(z) = \frac{1}{2}\pi(\sin(\pi\nu))^{-1}[I_{-\nu}(z) - I_\nu(z)] \quad \text{with} \quad I_\nu(z) = e^{-\frac{1}{2}\mathrm{i}\pi\nu} J_\nu(z\, e^{\frac{1}{2}\mathrm{i}\pi})$$ (modified Bessel functions, see 9.1.2.6, **3.**, p. 563),

$$C(x) = \frac{1}{\sqrt{2\pi}} \int_0^x \frac{\cos t}{\sqrt{t}}\, dt, \qquad S(x) = \frac{1}{\sqrt{2\pi}} \int_0^x \frac{\sin t}{\sqrt{t}}\, dt$$ (Fresnel integrals, see 14.4.3.2, **5.**, p. 757),

$$\operatorname{Si}(x) = \int_0^x \frac{\sin t}{t}\, dt, \qquad \operatorname{si}(x) = -\int_x^\infty \frac{\sin t}{t}\, dt = \operatorname{Si}(x) - \frac{\pi}{2}$$ (Integral sine, see 14.4.3.2, **2.**, p. 756),

$$\operatorname{Ci}(x) = -\int_x^\infty \frac{\cos t}{t}\, dt$$ (Integral cosine, see 14.4.3.2, **2.**, p. 756).

The abbreviations for functions occurring in the table correspond to those introduced in the corresponding chapters.

21.14.1 Fourier Cosine Transformation

No.	$f(t)$		$F_c(\omega) = \int_0^\infty f(t)\cos(t\,\omega)\,dt$
1.	$1,$ $0,$	$0 < t < a$ $t > a$	$\dfrac{\sin(a\,\omega)}{\omega}$
2.	$t,$ $2-t,$ $0,$	$0 < t < 1$ $1 < t < 2$ $t > 2$	$4\left(\cos\omega \sin^2\dfrac{\omega}{2}\right)\omega^{-2}$
3.	$0,$ $\dfrac{1}{t},$	$0 < t < a$ $t > a$	$-\operatorname{Ci}(a\,\omega)$
4.	$\dfrac{1}{\sqrt{t}}$		$\sqrt{\dfrac{\pi}{2}}\,\dfrac{1}{\sqrt{\omega}}$
5.	$\dfrac{1}{\sqrt{t}},$ $0,$	$0 < t < a$ $t > a$	$\sqrt{\dfrac{\pi}{2}}\,\dfrac{2\,C\,(a\omega)}{\sqrt{\omega}}$

No.	$f(t)$		$F_c(\omega) = \int\limits_0^\infty f(t)\,\cos(t\,\omega)\,dt$
6.	$0\,,$ $\dfrac{1}{\sqrt{t}}\,,$	$0 < t < a$ $t > a$	$\sqrt{\dfrac{\pi}{2}}\,\dfrac{1 - 2\,C\,(a\omega)}{\sqrt{\omega}}$
7.	$(a+t)^{-1}\,,$	$a > 0$	$[-\operatorname{si}(a\omega)\sin(a\omega) - \operatorname{Ci}(a\omega)\cos(a\omega)]$
8.	$(a-t)^{-1}\,,$	$a > 0$	$\left[\cos(a\omega)\operatorname{Ci}(a\omega) + \sin(a\omega)\left(\dfrac{\pi}{2} + \operatorname{Si}(a\omega)\right)\right]$
9.	$(a^2+t^2)^{-1}$		$\dfrac{\pi}{2}\,\dfrac{e^{-a\omega}}{a}$
10.	$(a^2-t^2)^{-1}$		$\dfrac{\pi}{2}\,\dfrac{\sin(a\omega)}{\omega}$
11.	$\dfrac{b}{b^2+(a-t)^2} + \dfrac{b}{b^2+(a+t)^2}$		$\pi\,e^{-b\omega}\cos(a\omega)$
12.	$\dfrac{a+t}{b^2+(a+t)^2} + \dfrac{a-t}{b^2+(a-t)^2}$		$\pi\,e^{-b\omega}\,\sin(a\omega)$
13.	$(a^2+t^2)^{-\frac{1}{2}}$		$K_0\,(a\omega)$
14.	$(a^2-t^2)^{-\frac{1}{2}}\,,$ $0\,,$	$0 < t < a$ $t > a$	$\dfrac{\pi}{2}\,J_0\,(a\omega)$
15.	$t^{-\nu}\,,$	$0 < \operatorname{Re}\nu < 1$	$\sin\left(\dfrac{\pi\,\nu}{2}\right)\,\Gamma\,(1-\nu)\,\omega^{\nu-1}$
16.	$e^{-a\,t}$		$\dfrac{a}{a^2+\omega^2}$
17.	$\dfrac{e^{-bt} - e^{-at}}{t}$		$\dfrac{1}{2}\,\ln\left(\dfrac{a^2+\omega^2}{b^2+\omega^2}\right)$
18.	$\sqrt{t}\,e^{-at}$		$\dfrac{\sqrt{\pi}}{2}\,(a^2+\omega^2)^{-\frac{3}{4}}\cos\left(\dfrac{3}{2}\arctan\left(\dfrac{\omega}{a}\right)\right)$
19.	$\dfrac{e^{-at}}{\sqrt{t}}$		$\sqrt{\dfrac{\pi}{2}}\left(\dfrac{a + (a^2+\omega^2)^{\frac{1}{2}}}{a^2+\omega^2}\right)^{\frac{1}{2}}$

No.	$f(t)$	$F_c(\omega) = \int_0^\infty f(t)\cos(t\,\omega)\,dt$
20.	$t^n e^{-at}$	$n!\,a^{n+1}(a^2+\omega^2)^{-(n+1)} \sum_{0\le 2m\le n+1} (-1)^m \binom{n+1}{2m}\left(\frac{\omega}{a}\right)^{2m}$
21.	$t^{\nu-1}\,e^{-at}$	$\Gamma(\nu)\,(a^2+\omega^2)^{-\frac{\nu}{2}}\cos\left(\nu\arctan\left(\frac{\omega}{a}\right)\right)$
22.	$\frac{1}{t}\left(\frac{1}{2}-\frac{1}{t}+\frac{1}{e^t-1}\right)$	$-\frac{1}{2}\ln(1-e^{-2\pi\omega})$
23.	$e^{-a\,t^2}$	$\frac{\sqrt{\pi}}{2}a^{-\frac{1}{2}}e^{-\frac{\omega^2}{4a}}$
24.	$t^{-\frac{1}{2}}e^{-\frac{a}{t}}$	$\sqrt{\frac{\pi}{2}}\,\frac{1}{\sqrt{\omega}}e^{-\sqrt{2a\omega}}(\cos\sqrt{2a\omega}-\sin\sqrt{2a\omega})$
25.	$t^{-\frac{3}{2}}e^{-\frac{a}{t}}$	$\sqrt{\frac{\pi}{a}}\,e^{-\sqrt{2a\omega}}\cos\sqrt{2a\omega}$
26.	$\ln t\,, \quad 0<t<1$ $0\,, \quad t>1$	$-\frac{\mathrm{Si}\,(\omega)}{\omega}$
27.	$\frac{\ln t}{\sqrt{t}}$	$-\sqrt{\frac{\pi}{2\omega}}\left(C+\frac{\pi}{2}+\ln 4\omega\right)$
28.	$(t^2-a^2)^{-1}\ln\left(\frac{t}{a}\right)$	$\frac{\pi}{2}\,\frac{1}{a}(\sin(a\omega)\ \mathrm{Ci}\,(a\omega)-\cos(a\omega)\,\mathrm{si}\,(a\omega))$
29.	$(t^2-a^2)^{-1}\ln(bt)$	$\frac{\pi}{2}\,\frac{1}{a}\ \{\sin(a\omega)\ [\mathrm{Ci}\,(a\omega)-\ln(ab)]-\cos(a\omega)\,\mathrm{si}(a\omega)\}$
30.	$\frac{1}{t}\ \ln(1+t)$	$\frac{1}{2}\left[\left(\mathrm{Ci}\left(\frac{\omega}{2}\right)\right)^2+\left(\mathrm{si}\left(\frac{\omega}{2}\right)\right)^2\right]$
31.	$\ln\left\|\frac{a+t}{b-t}\right\|$	$\frac{1}{\omega}\left\{\frac{\pi}{2}\ [\cos(b\omega)-\cos(a\omega)] + \cos(b\omega)\,\mathrm{Si}\,(b\omega)+\cos(a\omega)\,\mathrm{Si}\,(a\omega) - \sin(a\omega)\ \mathrm{Ci}\,(a\omega)-\sin(b\omega)\ \mathrm{Ci}\,(b\omega)\right\}$
32.	$e^{-at}\ \ln t$	$-\frac{1}{a^2+\omega^2}\left[aC+\frac{a}{2}\ \ln(a^2+\omega^2)+\omega\ \arctan\left(\frac{\omega}{a}\right)\right]$

No.	$f(t)$	$F_c(\omega) = \int_0^\infty f(t)\cos(t\omega)\,dt$
33.	$\ln\left(\dfrac{a^2+t^2}{b^2+t^2}\right)$	$\dfrac{\pi}{\omega}\left(e^{-b\omega}-e^{-a\omega}\right)$
34.	$\ln\left\vert\dfrac{a^2+t^2}{b^2-t^2}\right\vert$	$\dfrac{\pi}{\omega}\left(\cos(b\omega)-e^{-a\omega}\right)$
35.	$\dfrac{1}{t}\ln\left(\dfrac{a+t}{a-t}\right)^2$	$-2\pi\,\mathrm{si}(a\omega)$
36.	$\dfrac{\ln(a^2+t^2)}{\sqrt{a^2+t^2}}$	$-\left[\left(C+\ln\left(\dfrac{2\omega}{a}\right)\right)K_0(a\omega)\right]$
37.	$\ln\left(1+\dfrac{a^2}{t^2}\right)$	$\pi\,\dfrac{1-e^{-a\omega}}{\omega}$
38.	$\ln\left\vert 1-\dfrac{a^2}{t^2}\right\vert$	$\pi\,\dfrac{1-\cos(a\omega)}{\omega}$
39.	$\dfrac{\sin(at)}{t}$	$\dfrac{\pi}{2}, \quad \omega<a$ $\dfrac{\pi}{4}, \quad \omega=a$ $0, \quad \omega>a$
40.	$\dfrac{t\sin(at)}{t^2+b^2}$	$\dfrac{\pi}{2}e^{-ab}\cosh(b\omega), \quad \omega<a$ $-\dfrac{\pi}{2}e^{-b\omega}\sinh(ab), \quad \omega>a$
41.	$\dfrac{\sin(at)}{t(t^2+b^2)}$	$\dfrac{\pi}{2}b^{-2}\left(1-e^{-ab}\cosh(b\omega)\right), \quad \omega<a$ $\dfrac{\pi}{2}b^{-2}e^{-b\omega}\sinh(ab), \quad \omega>a$
42.	$e^{-bt}\sin(at)$	$\dfrac{1}{2}\left[\dfrac{a+\omega}{b^2+(a+\omega)^2}+\dfrac{a-\omega}{b^2+(a-\omega)^2}\right]$
43.	$\dfrac{e^{-t}\sin t}{t}$	$\dfrac{1}{2}\arctan\left(\dfrac{2}{\omega^2}\right)$
44.	$\dfrac{\sin^2(at)}{t}$	$\dfrac{1}{4}\ln\left\vert 1-4\dfrac{a^2}{\omega^2}\right\vert$
45.	$\dfrac{\sin(at)\sin(bt)}{t}$	$\dfrac{1}{2}\ln\left\vert\dfrac{(a+b)^2-\omega^2}{(a-b)^2-\omega^2}\right\vert$

No.	$f(t)$	$F_c(\omega) = \int_0^\infty f(t)\cos(t\omega)\,dt$
46.	$\dfrac{\sin^2(at)}{t^2}$	$\dfrac{\pi}{2}\left(a - \dfrac{1}{2}\omega\right), \quad \omega < 2a$ $0, \quad \omega > 2a$
47.	$\dfrac{\sin^3(at)}{t^2}$	$\dfrac{1}{8}\Big\{(\omega+3a)\,\ln(\omega+3a)$ $+(\omega-3a)\,\ln\lvert\omega-3a\rvert - (\omega+a)\,\ln(\omega+a)$ $-(\omega-a)\,\ln\lvert\omega-a\rvert\Big\}$
48.	$\dfrac{\sin^3(at)}{t^3}$	$\dfrac{\pi}{8}(3a^2-\omega^2), \quad 0 < \omega < a$ $\dfrac{\pi}{4}\omega^2, \quad \omega = a$ $\dfrac{\pi}{16}(3a-\omega)^2, \quad a < \omega < 3a$ $0, \quad \omega > 3a$
49.	$\dfrac{1-\cos(at)}{t}$	$\dfrac{1}{2}\,\ln\left\lvert 1 - \dfrac{a^2}{\omega^2}\right\rvert$
50.	$\dfrac{1-\cos(at)}{t^2}$	$\dfrac{\pi}{2}(a-\omega), \quad \omega < a$ $0, \quad \omega > a$
51.	$\dfrac{\cos(at)}{b^2+t^2}$	$\dfrac{\pi}{2}\,\dfrac{e^{-ab}\cosh(b\omega)}{b}, \quad \omega < a$ $\dfrac{\pi}{2}\,\dfrac{e^{-b\omega}\cosh(ab)}{b}, \quad \omega > a$
52.	$e^{-bt}\cos(at)$	$\dfrac{b}{2}\left[\dfrac{1}{b^2+(a-\omega)^2} + \dfrac{1}{b^2+(a+\omega)^2}\right]$
53.	$e^{-bt^2}\cos(at)$	$\dfrac{1}{2}\sqrt{\dfrac{\pi}{b}}\,e^{-\dfrac{a^2+\omega^2}{4b}}\cosh\left(\dfrac{a\omega}{2b}\right)$
54.	$\dfrac{t}{b^2+t^2}\tan(at)$	$\pi\cosh(b\omega)\,(1+e^{2ab})^{-1}$
55.	$\dfrac{t}{b^2+t^2}\cot(at)$	$\pi\cosh(b\omega)\,(e^{2ab}-1)^{-1}$

No.	$f(t)$	$F_c(\omega) = \int_0^\infty f(t)\cos(t\,\omega)\,dt$
56.	$\sin(at^2)$	$\frac{1}{2}\sqrt{\frac{\pi}{2a}}\left(\cos\left(\frac{\omega^2}{4a}\right) - \sin\left(\frac{\omega^2}{4a}\right)\right)$
57.	$\sin[a(1-t^2)]$	$-\frac{1}{2}\sqrt{\frac{\pi}{a}}\cos\left(a + \frac{\pi}{4} + \frac{\omega^2}{4a}\right)$
58.	$\frac{\sin(at^2)}{t^2}$	$\frac{\pi}{2}\,\omega\left[S\left(\frac{\omega^2}{4a}\right) - C\left(\frac{\omega^2}{4a}\right)\right] + \sqrt{2a}\sin\left(\frac{\pi}{4} + \frac{\omega^2}{4a}\right)$
59.	$\frac{\sin(at^2)}{t}$	$\frac{\pi}{2}\left\{\frac{1}{2} - \left[C\left(\frac{\omega^2}{4a}\right)\right]^2 - \left[S\left(\frac{\omega^2}{4a}\right)\right]^2\right\}$
60.	$e^{-at^2}\sin(bt^2)$	$\frac{\sqrt{\pi}}{2}(a^2+b^2)^{-\frac{1}{4}}\,e^{-\frac{1}{4}a\omega^2(a^2+b^2)^{-1}}$ $\cdot\sin\left[\frac{1}{2}\arctan\left(\frac{b}{a}\right) - \frac{b\omega^2}{4(a^2+b^2)}\right]$
61.	$\cos(at^2)$	$\frac{1}{2}\sqrt{\frac{\pi}{2a}}\left[\cos\left(\frac{\omega^2}{4a}\right) + \sin\left(\frac{\omega^2}{4a}\right)\right]$
62.	$\cos[a(1-t^2)]$	$\frac{1}{2}\sqrt{\frac{\pi}{a}}\sin\left(a + \frac{\pi}{4} + \frac{\omega^2}{4a}\right)$
63.	$e^{-at^2}\cos(bt^2)$	$\frac{\sqrt{\pi}}{2}(a^2+b^2)^{-\frac{1}{4}}\,e^{-\frac{1}{4}a\omega^2(a^2+b^2)^{-1}}$ $\cdot\cos\left[\frac{b\omega^2}{4(a^2+b^2)} - \frac{1}{2}\arctan\left(\frac{b}{a}\right)\right]$
64.	$\frac{1}{t}\sin\left(\frac{a}{t}\right)$	$\frac{\pi}{2}J_0(2\sqrt{a\omega})$
65.	$\frac{1}{\sqrt{t}}\sin\left(\frac{a}{t}\right)$	$\frac{1}{2}\sqrt{\frac{\pi}{2\omega}}\left[\sin(2\sqrt{a\omega}) + \cos(2\sqrt{a\omega}) - e^{-2\sqrt{a\omega}}\right]$
66.	$\left(\frac{1}{\sqrt{t}}\right)^3\sin\left(\frac{a}{t}\right)$	$\frac{1}{2}\sqrt{\frac{\pi}{2a}}\left[\sin(2\sqrt{a\omega}) + \cos(2\sqrt{a\omega}) + e^{-2\sqrt{a\omega}}\right]$
67.	$\frac{1}{\sqrt{t}}\cos\left(\frac{a}{t}\right)$	$\frac{1}{2}\sqrt{\frac{\pi}{2\omega}}\left[\cos(2\sqrt{a\omega}) - \sin(2\sqrt{a\omega}) + e^{-2\sqrt{a\omega}}\right]$
68.	$\left(\frac{1}{\sqrt{t}}\right)^3\cos\left(\frac{a}{t}\right)$	$\frac{1}{2}\sqrt{\frac{\pi}{2a}}\left[\cos(2\sqrt{a\omega}) - \sin(2\sqrt{a\omega}) + e^{-2\sqrt{a\omega}}\right]$

No.	$f(t)$	$F_c(\omega) \;=\; \int\limits_0^\infty f(t)\,\cos(t\,\omega)\,dt$
69.	$\frac{1}{\sqrt{t}}\,\sin\left(a\sqrt{t}\right)$	$2\sqrt{\frac{\pi}{2\omega}}\left[C\left(\frac{a^2}{4\omega}\right)\,\sin\left(\frac{a^2}{4\omega}\right) - S\left(\frac{a^2}{4\omega}\right)\,\cos\left(\frac{a^2}{4\omega}\right)\right]$
70.	$e^{-bt}\,\sin\left(a\sqrt{t}\right)$	$\frac{a}{2}\sqrt{\pi}\,(a^2+b^2)^{\frac{3}{4}}\,e^{-\frac{1}{4}\,a^2 b\,(b^2+\omega^2)^{-1}}$ $\cdot x\ cos\left[\frac{a^2\omega}{4\,(b^2+\omega^2)} - \frac{3}{2}\,\arctan\left(\frac{\omega}{b}\right)\right]$
71.	$\frac{\sin\left(a\sqrt{t}\right)}{t}$	$\pi\left[S\left(\frac{a^2}{4\omega}\right) + C\left(\frac{a^2}{4\omega}\right)\right]$
72.	$\frac{1}{\sqrt{t}}\,\cos\left(a\sqrt{t}\right)$	$\sqrt{\frac{\pi}{\omega}}\,\sin\left(\frac{\pi}{4} + \frac{a^2}{4\omega}\right)$
73.	$\frac{e^{-at}}{\sqrt{t}}\,\cos\left(b\sqrt{t}\right)$	$\sqrt{\pi}\,(a^2+\omega^2)^{-\frac{1}{4}}\,e^{-\frac{1}{4}\,ab^2\,(a^2+b^2)^{-1}}$ $\cdot\,\cos\left[\frac{b^2\omega}{4\,(a^2+\omega^2)} - \frac{1}{2}\,\arctan\left(\frac{\omega}{a}\right)\right]$
74.	$e^{-a\sqrt{t}}\,\cos\left(a\sqrt{t}\right)$	$\sqrt{\pi}\,a\,(2\omega)^{-\frac{3}{2}}\,e^{-\frac{a^2}{2\omega}}$
75.	$\frac{e^{-a\sqrt{t}}}{\sqrt{t}}\left[\cos\left(a\sqrt{t}\right) - \sin\left(a\sqrt{t}\right)\right]$	$\sqrt{\frac{\pi}{2\omega}}\,e^{-\frac{a^2}{2\omega}}$

21.14.2 Fourier Sine Transformation

No.	$f(t)$	$F_s(\omega) \;=\; \int\limits_0^\infty f(t)\,\sin(t\,\omega)\,dt$
1.	$1,\quad 0<t<a$ $0,\quad t>a$	$\frac{1-\cos(a\omega)}{\omega}$
2.	$t,\quad 0<t<1$ $2-t,\quad 1<t<2$ $0,\quad t>2$	$4\omega^{-2}\,\sin\omega\,\sin^2\left(\frac{\omega}{2}\right)$
3.	$\frac{1}{t}$	$\frac{\pi}{2}$
4.	$\frac{1}{t},\quad 0<t<a$ $0,\quad t>a$	$\mathrm{Si}\,(a\omega)$

No.	$f(t)$	$F_s(\omega) = \int_0^\infty f(t)\,\sin(t\,\omega)\,dt$
5.	$0\,, \quad 0 < t < a$ $\frac{1}{t}\,, \quad t > a$	$-\,\mathrm{si}\,(a\omega)$
6.	$\frac{1}{\sqrt{t}}$	$\sqrt{\frac{\pi}{2}}\,\frac{1}{\sqrt{\omega}}$
7.	$\frac{1}{\sqrt{t}}\,, \quad 0 < t < a$ $0, \quad t > a$	$\sqrt{\frac{\pi}{2}}\,\frac{2\,S(a\omega)}{\sqrt{\omega}}$
8.	$0, \quad 0 < t < a$ $\frac{1}{\sqrt{t}}, \quad t > a$	$\sqrt{\frac{\pi}{2}}\,\frac{1 - 2\,S(a\omega)}{\sqrt{\omega}}$
9.	$\left(\frac{1}{\sqrt{t}}\right)^3$	$\sqrt{\pi\,2\,\omega}$
10.	$(a+t)^{-1} \qquad (a > 0)$	$[\sin\,(a\omega)\;\mathrm{Ci}\,(a\omega) - \cos\,(a\omega)\;\mathrm{si}\,(a\omega)]$
11.	$(a-t)^{-1} \qquad (a > 0)$	$\left[\sin\,(a\omega)\;\mathrm{Ci}(a\omega) - \cos\,(a\omega)\;\left(\frac{\pi}{2} + \mathrm{Si}(a\omega)\right)\right]$
12.	$\frac{t}{a^2+t^2}$	$\frac{\pi}{2}\,e^{-\,a\omega}$
13.	$(a^2-t^2)^{-1}$	$\frac{1}{a}\,[\sin\,(a\omega)\;\mathrm{Ci}(a\omega) - \cos\,(a\omega)\,\mathrm{Si}(a\omega))$
14.	$\frac{b}{b^2+(a-t)^2} - \frac{b}{b^2+(a+t)^2}$	$\pi\,e^{-b\omega}\sin\,(a\omega)$
15.	$\frac{a+t}{b^2+(a+t)^2} - \frac{a-t}{b^2+(a-t)^2}$	$\pi\,e^{-b\omega}\cos\,(a\omega)$
16.	$\frac{t}{a^2-t^2}$	$-\frac{\pi}{2}\,\cos\,(a\omega)$
17.	$\frac{1}{t\,(a^2-t^2)}$	$\frac{\pi}{2}\,\frac{1-\cos\,(a\omega)}{a^2}$
18.	$\frac{1}{t\,(a^2+t^2)}$	$\frac{\pi}{2}\,\frac{1-e^{-\,a\omega}}{a^2}$

No.	$f(t)$	$F_s(\omega) = \int_0^\infty f(t)\,\sin(t\,\omega)\,dt$
19.	$t^{-\nu}, \qquad 0 < \mathrm{Re}\,\nu < 2$	$\cos\left(\frac{\pi\,\nu}{2}\right)\,\Gamma\,(1-\nu)\,\omega^{\nu-1}$
20.	e^{-at}	$\frac{\omega}{a^2+\omega^2}$
21.	$\frac{e^{-at}}{t}$	$\arctan\left(\frac{\omega}{a}\right)$
22.	$\frac{e^{-at}-e^{-bt}}{t^2}$	$\left[\frac{1}{2}\,\omega\,\ln\left(\frac{b^2+\omega^2}{a^2+\omega^2}\right)+b\,\arctan\left(\frac{\omega}{b}\right)-a\,\arctan\left(\frac{\omega}{a}\right)\right]$
23.	$\sqrt{t}\,e^{-at}$	$\frac{\sqrt{\pi}}{2}\,(a^2+\omega^2)^{-\frac{3}{4}}\,\sin\left[\frac{3}{2}\,\arctan\left(\frac{\omega}{a}\right)\right]$
24.	$\frac{e^{-at}}{\sqrt{t}}$	$\left(\frac{(a^2+\omega^2)^{\frac{1}{2}}-a}{a^2+\omega^2}\right)^{\frac{1}{2}}$
25.	$t^n\,e^{-at}$	$n\,!a^{n+1}(a^2+\omega^2)^{-(n+1)}\quad\sum_{m=0}^{[\frac{1}{2}n]}(-1)^m\binom{n+1}{2m+1}\left(\frac{\omega}{a}\right)^{2m+1}$
26.	$t^{\nu-1}\,e^{-at}$	$\Gamma(\nu)\,(a^2+\omega^2)^{-\frac{\nu}{2}}\,\sin\left[\nu\,\arctan\left(\frac{\omega}{a}\right)\right]$
27.	$e^{-\frac{1}{2}t}\,(1-e^{-t})^{-1}$	$-\frac{1}{2}\,\tanh\,(\pi\omega)$
28.	$t\,e^{-at^2}$	$\sqrt{\frac{\pi}{a}}\,\frac{\omega}{4a}\,e^{-\frac{\omega^2}{4a}}$
29.	$t^{-\frac{1}{2}}\,e^{-\frac{a}{t}}$	$\sqrt{\frac{\pi}{2\omega}}\,e^{-\sqrt{2a\omega}}\,[\cos\,\sqrt{2a\omega}+\sin\,\sqrt{2a\omega}]$
30.	$t^{-\frac{3}{2}}\,e^{-\frac{a}{t}}$	$\sqrt{\frac{\pi}{\omega}}\,e^{-\sqrt{2a\omega}}\,\sin\,\sqrt{2a\omega}$
31.	$\ln\,t, \qquad 0 < t < 1$ $0, \qquad t > 1$	$\frac{\mathrm{Ci}(\omega)-C-\ln\,\omega}{\omega}$

No.	$f(t)$	$F_s(\omega) = \int\limits_0^\infty f(t)\, \sin(t\,\omega)\, dt$
32.	$\dfrac{\ln t}{t}$	$-\dfrac{\pi}{2}\,(C + \ln\omega)$
33.	$\dfrac{\ln t}{\sqrt{t}}$	$\sqrt{\dfrac{\pi}{2\omega}}\,\left[\dfrac{\pi}{2} - C - \ln 4\omega\right]$
34.	$t\,(t^2 - a^2)^{-1}\,\ln(bt)$	$\dfrac{\pi}{2}\,[\cos(a\omega)\,(\ln(ab) - \mathrm{Ci}(a\omega)) - \sin(a\omega)\cdot \mathrm{si}(a\omega)]$
35.	$t\,(t^2 - a^2)^{-1}\,\ln(\dfrac{t}{a})$	$-\dfrac{\pi}{2}\,[\cos(a\omega)\;\mathrm{Ci}(a\omega) + \sin(a\omega)\,si\,(a\omega)]$
36.	$e^{-at}\,\ln t$	$\dfrac{1}{a^2+\omega^2}\,\left[a\,\arctan\left(\dfrac{\omega}{a}\right) - C\omega - \dfrac{1}{2}\,\omega\,\ln(a^2+\omega^2)\right]$
37.	$\ln\left\|\dfrac{a+t}{b-t}\right\|$	$\dfrac{1}{\omega}\,\Big\{\ln\left(\dfrac{a}{b}\right) + \cos(b\omega)\;\mathrm{Ci}(b\omega) - \cos(a\omega)\;\mathrm{Ci}(a\omega)$ $+ \sin(b\omega)\,\mathrm{Si}(b\omega) - \sin(a\omega)\,\mathrm{Si}(a\omega)$ $+\dfrac{\pi}{2}\,[\sin(b\omega) + \sin(a\omega)]\Big\}$
38.	$\ln\left\|\dfrac{a+t}{a-t}\right\|$	$\dfrac{\pi}{\omega}\,\sin(a\omega)$
39.	$\dfrac{1}{t^2}\,\ln\left(\dfrac{a+t}{a-t}\right)^2$	$\dfrac{2\pi}{a}\,[1 - \cos(a\omega) - a\omega\;\mathrm{si}(a\omega)]$
40.	$\ln\left(\dfrac{a^2+t^2+t}{a^2+t^2-t}\right)$	$\dfrac{2\pi}{\omega}\,e^{-\omega\sqrt{a^2-\frac{1}{4}}}\,\sin\left(\dfrac{\omega}{2}\right)$
41.	$\ln\left\|1 - \dfrac{a^2}{t^2}\right\|$	$\dfrac{2}{\omega}\,[C + \ln(a\omega) - \cos(a\omega)\;\mathrm{Ci}(a\omega) - \sin(a\omega)\,\mathrm{Si}(a\omega)]$
42.	$\ln\left(\dfrac{a^2+(b+t)^2}{a^2+(b-t)^2}\right)$	$\dfrac{2\pi}{\omega}\,e^{-a\omega}\,\sin(b\omega)$
43.	$\dfrac{1}{t}\,\ln\|1 - a^2t^2\|$	$-\,\pi\;\mathrm{Ci}\left(\dfrac{\omega}{a}\right)$
44.	$\dfrac{1}{t}\,\ln\left\|1 - \dfrac{a^2}{t^2}\right\|$	$\pi\,[C + \ln(a\omega) - \mathrm{Ci}(a\omega)]$

No.	$f(t)$	$F_s(\omega) = \int_0^\infty f(t)\,\sin(t\,\omega)\,dt$
45.	$\dfrac{\sin(at)}{t}$	$\dfrac{1}{2}\ln\left\vert\dfrac{\omega+a}{\omega-a}\right\vert$
46.	$\dfrac{\sin(at)}{t^2}$	$\dfrac{\pi}{2}\omega, \quad 0<\omega<a$ $\dfrac{\pi}{2}a, \quad \omega>a$
47.	$\dfrac{\sin(\pi t)}{1-t^2}$	$\sin\omega, \quad 0\le\omega\le\pi$ $0, \quad \omega\ge\pi$
48.	$\dfrac{\sin(at)}{b^2+t^2}$	$\dfrac{\pi}{2}\dfrac{e^{-ab}}{b}\sinh(b\omega), \quad 0<\omega<a$ $\dfrac{\pi}{2}\dfrac{e^{-b\omega}}{b}\sinh(ab), \quad \omega>a$
49.	$e^{-bt}\sin(at)$	$\dfrac{1}{2}b\left[\dfrac{1}{b^2+(a-\omega)^2}-\dfrac{1}{b^2+(a+\omega)^2}\right]$
50.	$\dfrac{e^{-bt}\sin(at)}{t}$	$\dfrac{1}{4}\ln\left(\dfrac{b^2+(\omega+a)^2}{b^2+(\omega-a)^2}\right)$
51.	$e^{-bt^2}\sin(at)$	$\dfrac{1}{2}\sqrt{\dfrac{\pi}{b}}\,e^{-\frac{1}{4}\frac{a^2+\omega^2}{b}}\sinh\left(\dfrac{a\omega}{2b}\right)$
52.	$\dfrac{\sin^2(at)}{t}$	$\dfrac{\pi}{4}, \quad 0<\omega<2a$ $\dfrac{\pi}{8}, \quad \omega=2a$ $0, \quad \omega>2a$
53.	$\dfrac{\sin(at)\sin(bt)}{t}$	$0, \quad 0<\omega<a-b$ $\dfrac{\pi}{4}, \quad a-b<\omega<a+b$ $0, \quad \omega>a+b$
54.	$\dfrac{\sin^2(at)}{t^2}$	$\dfrac{1}{4}\Big[(\omega+2a)\ln(\omega+2a)$ $+(\omega-2a)\ln\vert\omega-2a\vert-\dfrac{1}{2}\omega\ln\omega\Big]$

No.	$f(t)$	$F_s(\omega) = \int_0^\infty f(t)\,\sin(t\,\omega)\,dt$
55.	$\frac{\sin^2(at)}{t^3}$	$\frac{\pi}{4}\omega\left(2a-\frac{\omega}{2}\right), \quad 0<\omega<2a$ $\frac{\pi}{2}a^2, \quad \omega>2a$
56.	$\frac{\cos(at)}{t}$	$0, \quad 0<\omega<a$ $\frac{\pi}{4}, \quad \omega=a$ $\frac{\pi}{2}, \quad \omega>a$
57.	$\frac{t\,\cos(at)}{b^2+t^2}$	$-\frac{\pi}{2}e^{-ab}\sinh(b\omega), \quad 0<\omega<a$ $\frac{\pi}{2}e^{-b\omega}\cosh(ab), \quad \omega>a$
58.	$\sin(at^2)$	$\sqrt{\frac{\pi}{2a}}\left[\cos\left(\frac{\omega^2}{4a}\right)C\left(\frac{\omega^2}{4a}\right)+\sin\left(\frac{\omega^2}{4a}\right)S\left(\frac{\omega^2}{4a}\right)\right]$
59.	$\frac{\sin(at^2)}{t}$	$\frac{\pi}{2}\left[C\left(\frac{\omega^2}{4a}\right)-S\left(\frac{\omega^2}{4a}\right)\right]$
60.	$\cos(at^2)$	$\sqrt{\frac{\pi}{2a}}\left[\sin\left(\frac{\omega^2}{4a}\right)C\left(\frac{\omega^2}{4a}\right)-\cos\left(\frac{\omega^2}{4a}\right)S\left(\frac{\omega^2}{4a}\right)\right]$
61.	$\frac{\cos(at^2)}{t}$	$\frac{\pi}{2}\left[C\left(\frac{\omega^2}{4a}\right)+S\left(\frac{\omega^2}{4a}\right)\right]$
62.	$e^{-a\sqrt{t}}\sin(a\sqrt{t})$	$\sqrt{\frac{\pi}{2}}\frac{a}{2\omega\sqrt{\omega}}e^{-\frac{a^2}{2\omega}}$

21.14.3 Fourier Transformation

Although $F(\omega)$ can be represented by the Fourier cosine transformation F_c and the Fourier sine transformation F_s according to (15.75a), here we give some direct transforms $F(\omega)$.

No.	$f(t)$	$F(\omega) = \int_{-\infty}^{\infty} e^{-i\omega t}f(t)\,dt$
1.	$\delta(t)$ (Dirac δ function)	1
2.	$\delta^{(n)}(t)$	$(i\,\omega)^n$
3.	$\delta^{(n)}(t-a)$	$(i\,\omega)^n e^{-ia\omega} \quad (n=0,1,2,\ldots)$

No.	$f(t)$	$F(\omega) = \int\limits_{-\infty}^{\infty} e^{-i\omega t} f(t)\, dt$
4.	1	$2\pi\delta(\omega)$
5.	t^n	$2\pi \mathrm{i}^n\, \delta^{(n)}(\omega) \quad (n = 1, 2, \ldots)$
6.	$H(t) = 1$ for $t > 0$ $H(t) = 0$ for $t < 0$ (Heaviside unit step function)	$\frac{1}{\mathrm{i}\,\omega} + \pi\delta(\omega)$
7.	$t^n\, H(t)$	$\frac{n!}{(\mathrm{i}\,\omega)^{n+1}} + \pi \mathrm{i}^n \delta^{(n)}(\omega) \quad (n = 1, 2, \ldots)$
8.	$e^{-at}\, H(t) = e^{-at}$ for $t > 0$ $e^{-at}\, H(t) = 0$ for $t < 0$	$\frac{1}{a + \mathrm{i}\,\omega} \quad (a > 0)$
9.	$\frac{1}{\sqrt{4\pi a}} e^{-t^2/(4a)}$	$e^{-a\omega^2} \quad (a > 0)$
10.	$\frac{1}{2a} e^{-a\|t\|}$	$\frac{1}{\omega^2 + a^2} \quad (a > 0)$
11.	$\frac{1}{t^2 + a^2}$	$\frac{\pi}{a} e^{-a\|\omega\|}$
12.	$\frac{t}{t^2 + a^2}$	$-\mathrm{i}\pi e^{-a\|\omega\|} \operatorname{sign} \omega$
13.	$H(t + a) - H(t - a) = 1$ for $\|t\| < a$ $H(t + a) - H(t - a) = 0$ for $\|t\| > a$	$\frac{2 \sin a\,\omega}{\omega}$
14.	$e^{\mathrm{i}\,at}$	$2\pi\delta(\omega - a)$
15.	$\cos at$	$\pi[\delta(\omega + a) + \delta(\omega - a)]$
16.	$\sin at$	$\mathrm{i}\,\pi[\delta(\omega + a) - \delta(\omega - a)]$
17.	$\frac{1}{\cosh t}$	$\frac{\pi}{\cosh \frac{\pi\omega}{2}}$
18.	$\frac{1}{\sinh t}$	$-\mathrm{i}\pi \tanh \frac{\pi\omega}{2}$
19.	$\sin at^2$	$\sqrt{\frac{\pi}{a}} \left(\frac{\omega^2}{4a} + \frac{\pi}{4} \right) \quad (a > 0)$
20.	$\cos at^2$	$\sqrt{\frac{\pi}{a}} \left(\frac{\omega^2}{4a} - \frac{\pi}{4} \right) \quad (a > 0)$

21.14.4 Exponential Fourier Transformation

Although the exponential Fourier transformation $F_e(\omega)$ can be represented by the Fourier transformation $F(\omega)$ according to (15.77), i.e., $F_e(\omega) = \frac{1}{2}F(-\omega)$, here we give some direct transforms.

No.	$f(t)$	$F_e(\omega) = \frac{1}{2}\int\limits_{-\infty}^{\infty} f(t)\, e^{it\omega}\, dt$
1.	$f(t) = A$ for $a \le t \le b$ $f(t) = 0$ otherwise	$\frac{\mathrm{i}\,A}{2\omega}\,(e^{\mathrm{i}\,a\omega} - e^{\mathrm{i}\,b\omega})$
2.	$f(t) = t^n$ for $0 \le t \le b$ $f(t) = 0$ otherwise $(n = 1, 2, \ldots)$	$\frac{1}{2}\left[n!\,(-\mathrm{i}\,\omega)^{-(n+1)} - e^{\mathrm{i}\,b\omega}\sum\limits_{m=0}^{n}\frac{n!}{m!}\,(-\mathrm{i}\,\omega)^{m-n-1}\,b^m\right]$
3.	$\frac{1}{(a+\mathrm{i}\,t)^\nu}$ $\quad \mathrm{Re}\,\nu > 0$	$\frac{\pi}{\Gamma(\nu)}\,\omega^{\nu-1}\,e^{-a\omega}$ for $\omega > 0$ 0 for $\omega < 0$
4.	$\frac{1}{(a-\mathrm{i}\,t)^\nu}$ $\quad \mathrm{Re}\,\nu > 0$	0 for $\omega > 0$ $\frac{\pi}{\Gamma(\nu)}\,(-\omega)^{\nu-1}\,e^{a\omega}$ for $\omega < 0$

21.15 Z Transformation

For definition see 15.4.1.2, p. 794, for rules of calculations see 15.4.1.3, p. 795, for inverses see p. 797

No.	Original Sequence f_n	Transform $F(z) = Z(f_n)$	Convergence Region
1	1	$\frac{z}{z-1}$	$\lvert z\rvert > 1$
2	$(-1)^n$	$\frac{z}{z+1}$	$\lvert z\rvert > 1$
3	n	$\frac{z}{(z-1)^2}$	$\lvert z\rvert > 1$
4	n^2	$\frac{z(z+1)}{(z-1)^3}$	$\lvert z\rvert > 1$
5	n^3	$\frac{z(z^2+4z+1)}{(z-1)^4}$	$\lvert z\rvert > 1$
6	e^{an}	$\frac{z}{z-e^a}$	$\lvert z\rvert > \lvert e^a\rvert$
7	a^n	$\frac{z}{z-a}$	$\lvert z\rvert > \lvert a\rvert$
8	$\frac{a^n}{n!}$	$e^{\frac{a}{z}}$	$\lvert z\rvert > 0$
9	$n\,a^n$	$\frac{za}{(z-a)^2}$	$\lvert z\rvert > \lvert a\rvert$
10	$n^2\,a^n$	$\frac{az(z+a)}{(z-a)^3}$	$\lvert z\rvert > \lvert a\rvert$
11	$\binom{n}{k}$	$\frac{z}{(z-1)^{k+1}}$	$\lvert z\rvert > 1$
12	$\binom{k}{n}$	$\left(1+\frac{1}{z}\right)^k$	$\lvert z\rvert > 0$
13	$\sin bn$	$\frac{z\sin b}{z^2-2z\cos b+1}$	$\lvert z\rvert > 1$
14	$\cos bn$	$\frac{z(z-\cos b)}{z^2-2z\cos b+1}$	$\lvert z\rvert > 1$

No.	Original Sequence f_n	Transform $F(z) = Z(f_n)$	Convergence Region
15	$e^{an} \sin bn$	$\dfrac{ze^a \sin b}{z^2 - 2ze^a \cos b + e^{2a}}$	$\|z\| > \|e^a\|$
16	$e^{an} \cos bn$	$\dfrac{z(z - e^a \cos b)}{z^2 - 2ze^a \cos b + e^{2a}}$	$\|z\| > \|e^a\|$
17	$\sinh bn$	$\dfrac{z \sinh b}{z^2 - 2z \cosh b + 1}$	$\|z\| > \max(\|e^b\|, \|e^{-b}\|)$
18	$\cosh bn$	$\dfrac{z(z - \cosh b)}{z^2 - 2z \cosh b + 1}$	$\|z\| > \max(\|e^b\|, \|e^{-b}\|)$
19	$a^n \sinh bn$	$\dfrac{za \sinh b}{z^2 - 2za \cosh b + a^2}$	$\|z\| > \max(\|ae^b\|, \|ae^{-b}\|)$
20	$a^n \cosh bn$	$\dfrac{z(z - a \cosh b)}{z^2 - 2za \cosh b + a^2}$	$\|z\| > \max(\|ae^b\|, \|ae^{-b}\|)$
21	$f_n = 0 \quad f"ur \quad n \neq k,$ $f_k = 1$	$\dfrac{1}{z^k}$	$\|z\| > 0$
22	$f_{2n} = 0, \quad f_{2n+1} = 2$	$\dfrac{2z}{z^2 - 1}$	$\|z\| > 1$
23	$f_{2n} = 0,$ $f_{2n+1} = 2(2n+1)$	$\dfrac{2z(z^2+1)}{(z^2-1)^2}$	$\|z\| > 1$
24	$f_{2n} = 0,$ $f_{2n+1} = \dfrac{2}{2n+1}$	$\ln \dfrac{z-1}{z+1}$	$\|z\| > 1$
25	$\cos \dfrac{n\pi}{2}$	$\dfrac{z^2}{z^2+1}$	$\|z\| > 1$
26	$(n+1)\, e^{an}$	$\dfrac{z^2}{(z - e^a)^2}$	$\|z\| > \|e^a\|$
27	$\dfrac{e^{b(n+1)} - e^{a(n+1)}}{e^b - e^a}$	$\dfrac{z^2}{(z - e^a)\,(z - e^b)}$	$\|z\| > \max(\|e^a\|, \|e^b\|),\ a \neq b$
28	$\dfrac{1}{6}\,(n-1)\, n(n+1)$	$\dfrac{z^2}{(z-1)^4}$	$\|z\| > 1$

No.	**Original Sequence f_n**	**Transform $F(z) = Z(f_n)$**	**Convergence Region**
29	$f_0 = 0, \quad f_n = \frac{1}{n}, \quad n \geq 1$	$\ln \frac{z}{z-1}$	$\lvert z\rvert > 1$
30	$\frac{(-1)^n}{(2n+1)!}$	$\sqrt{z}\,\sin \frac{1}{\sqrt{z}}$	$\lvert z\rvert > 0$
31	$\frac{(-1)^n}{(2n)!}$	$\cos \frac{1}{\sqrt{z}}$	$\lvert z\rvert > 0$

21.16 Poisson Distribution

For the formula of the Poisson distribution see 16.2.3.3, p. 817.

k	λ					
	0.1	0.2	0.3	0.4	0.5	0.6
0	0.904837	0.818731	0.740818	0.670320	0.606531	0.548812
1	0.090484	0.163746	0.222245	0.268128	0.303265	0.329287
2	0.004524	0.016375	0.033337	0.053626	0.075816	0.098786
3	0.000151	0.001091	0.003334	0.007150	0.012636	0.019757
4	0.000004	0.000055	0.000250	0.000715	0.001580	0.002964
5		0.000002	0.000015	0.000057	0.000158	0.000356
6			0.000001	0.000004	0.000013	0.000035
7					0.000001	0.000003

k	λ					
	0.7	0.8	0.9	1.0	2.0	3.0
0	0.496585	0.449329	0.406570	0.367879	0.135335	0.049787
1	0.347610	0.359463	0.365913	0.367879	0.270671	0.149361
2	0.121663	0.143785	0.164661	0.183940	0.270671	0.224042
3	0.028388	0.038343	0.049398	0.061313	0.180447	0.224042
4	0.004968	0.007669	0.011115	0.015328	0.090224	0.168031
5	0.000696	0.001227	0.002001	0.003066	0.036089	0.100819
6	0.000081	0.000164	0.000300	0.000511	0.012030	0.050409
7	0.000008	0.000019	0.000039	0.000073	0.003437	0.021604
8	0.000001	0.000002	0.000004	0.000009	0.000859	0.008102
9				0.000001	0.000191	0.002701
10					0.000038	0.000810
11					0.000007	0.000221
12					0.000001	0.000055
13						0.000013
14						0.000003
15						0.000001

(continuation)

k	λ					
	4.0	5.0	6.0	7.0	8.0	9.0
0	0.018316	0.006738	0.002479	0.000912	0.000335	0.000123
1	0.073263	0.033690	0.014873	0.006383	0.002684	0.001111
2	0.146525	0.084224	0.044618	0.022341	0.010735	0.004998
3	0.195367	0.140374	0.089235	0.052129	0.028626	0.014994
4	0.195367	0.175467	0.133853	0.091126	0.057252	0.033737
5	0.156293	0.175467	0.160623	0.127717	0.091604	0.060727
6	0.104194	0.146223	0.160623	0.149003	0.122138	0.091090
7	0.059540	0.104445	0.137677	0.149003	0.139587	0.117116
8	0.029770	0.065278	0.103258	0.130377	0.139587	0.131756
9	0.013231	0.036266	0.068838	0.101405	0.124077	0.131756
10	0.005292	0.018133	0.041303	0.070983	0.099262	0.118580
11	0.001925	0.008242	0.022529	0.045171	0.072190	0.097020
12	0.000642	0.003434	0.011264	0.026350	0.048127	0.072765
13	0.000197	0.001321	0.005199	0.014188	0.029616	0.050376
14	0.000056	0.000472	0.002228	0.007094	0.016924	0.032384
15	0.000015	0.000157	0.000891	0.003311	0.009026	0.019431
16	0.000004	0.000049	0.000334	0.001448	0.004513	0.010930
17	0.000001	0.000014	0.000118	0.000596	0.002124	0.005786
18		0.000004	0.000039	0.000232	0.000944	0.002893
19		0.000001	0.000012	0.000085	0.000397	0.001370
20			0.000004	0.000030	0.000159	0.000617
21			0.000001	0.000010	0.000061	0.000264
22				0.000003	0.000022	0.000108
23				0.000001	0.000008	0.000042
24					0.000003	0.000016
25					0.000001	0.000006
26						0.000002
27						0.000001

21.17 Standard Normal Distribution

For the formula of the standard normal distribution see 16.2.4.2, p. 819.

21.17.1 Standard Normal Distribution for $0.00 \le x \le 1.99$

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
0.00	**0.5000**	**0.20**	**0.5793**	**0.40**	**0.6554**	**0.60**	**0.7257**	**0.80**	0.7881
0.01	0.5040	0.21	0.5832	0.41	0.6591	0.61	0.7291	0.81	0.7910
0.02	0.5080	0.22	0.5871	0.42	0.6628	0.62	0.7324	0.82	0.7939
0.03	0.5120	0.23	0.5910	0.43	0.6664	0.63	0.7357	0.83	0.7967
0.04	0.5160	0.24	0.5948	0.44	0.6700	0.64	0.7389	0.84	0.7995
0.05	0.5199	0.25	0.5987	0.45	0.6736	0.65	0.7422	0.85	0.8023
0.06	0.5239	0.26	0.6026	0.46	0.6772	0.66	0.7454	0.86	0.8051
0.07	0.5279	0.27	0.6064	0.47	0.6808	0.67	0.7486	0.87	0.8079
0.08	0.5319	0.28	0.6103	0.48	0.6844	0.68	0.7517	0.88	0.8106
0.09	0.5359	0.29	0.6141	0.49	0.6879	0.69	0.7549	0.89	0.8133
0.10	**0.5398**	**0.30**	**0.6179**	**0.50**	**0.6915**	**0.70**	**0.7580**	**0.90**	0.8159
0.11	0.5438	0.31	0.6217	0.51	0.6950	0.71	0.7611	0.91	0.8186
0.12	0.5478	0.32	0.6255	0.52	0.6985	0.72	0.7642	0.92	0.8212
0.13	0.5517	0.33	0.6293	0.53	0.7019	0.73	0.7673	0.93	0.8238
0.14	0.5557	0.34	0.6331	0.54	0.7054	0.74	0.7704	0.94	0.8264
0.15	0.5596	0.35	0.6368	0.55	0.7088	0.75	0.7734	0.95	0.8289
0.16	0.5636	0.36	0.6406	0.56	0.7123	0.76	0.7764	0.96	0.8315
0.17	0.5675	0.37	0.6443	0.57	0.7157	0.77	0.7794	0.97	0.8340
0.18	0.5714	0.38	0.6480	0.58	0.7190	0.78	0.7823	0.98	0.8365
0.19	0.5753	0.39	0.6517	0.59	0.7224	0.79	0.7852	0.99	0.8389

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
1.00	**0.8413**	**1.20**	**0.8849**	**1.40**	**0.9192**	**1.60**	**0.9452**	**1.80**	**0.9641**
1.01	0.8438	1.21	0.8869	1.41	0.9207	1.61	0.9463	1.81	0.9649
1.02	0.8461	1.22	0.8888	1.42	0.9222	1.62	0.9474	1.82	0.9656
1.03	0.8485	1.23	0.8907	1.43	0.9236	1.63	0.9484	1.83	0.9664
1.04	0.8508	1.24	0.8925	1.44	0.9251	1.64	0.9495	1.84	0.9671
1.05	0.8531	1.25	0.8944	1.45	0.9265	1.65	0.9505	1.85	0.9678
1.06	0.8554	1.26	0.8962	1.46	0.9279	1.66	0.9515	1.86	0.9686
1.07	0.8577	1.27	0.8980	1.47	0.9292	1.67	0.9525	1.87	0.9693
1.08	0.8599	1.28	0.8997	1.48	0.9306	1.68	0.9535	1.88	0.9699
1.09	0.8621	1.29	0.9015	1.49	0.9319	1.69	0.9545	1.89	0.9706
1.10	**0.8643**	**1.30**	**0.9032**	**1.50**	**0.9332**	**1.70**	**0.9554**	**1.90**	**0.9713**
1.11	0.8665	1.31	0.9049	1.51	0.9345	1.71	0.9564	1.91	0.9719
1.12	0.8686	1.32	0.9066	1.52	0.9357	1.72	0.9573	1.92	0.9726
1.13	0.8708	1.33	0.9082	1.53	0.9370	1.73	0.9582	1.93	0.9732
1.14	0.8729	1.34	0.9099	1.54	0.9382	1.74	0.9591	1.94	0.9738
1.15	0.8749	1.35	0.9115	1.55	0.9394	1.75	0.9599	1.95	0.9744
1.16	0.8770	1.36	0.9131	1.56	0.9406	1.76	0.9608	1.96	0.9750
1.17	0.8790	1.37	0.9147	1.57	0.9418	1.77	0.9616	1.97	0.9756
1.18	0.8810	1.38	0.9162	1.58	0.9429	1.78	0.9625	1.98	0.9761
1.19	0.8830	1.39	0.9177	1.59	0.9441	1.79	0.9633	1.99	0.9767

21.17.2 Standard Normal Distribution for $2.00 \le x \le 3.90$

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
2.00	**0.9773**	**2.20**	**0.9861**	**2.40**	**0.9918**	**2.60**	**0.9953**	**2.80**	**0.9974**
2.01	0.9778	2.21	0.9864	2.41	0.9920	2.61	0.9955	2.81	0.9975
2.02	0.9783	2.22	0.9868	2.42	0.9922	2.62	0.9956	2.82	0.9976
2.03	0.9788	2.23	0.9871	2.43	0.9925	2.63	0.9957	2.83	0.9977
2.04	0.9793	2.24	0.9875	2.44	0.9927	2.64	0.9959	2.84	0.9977
2.05	0.9798	2.25	0.9878	2.45	0.9929	2.65	0.9960	2.85	0.9978
2.06	0.9803	2.26	0.9881	2.46	0.9931	2.66	0.9961	2.86	0.9979
2.07	0.9808	2.27	0.9884	2.47	0.9932	2.67	0.9962	2.87	0.9979
2.08	0.9812	2.28	0.9887	2.48	0.9934	2.68	0.9963	2.88	0.9980
2.09	0.9817	2.29	0.9890	2.49	0.9936	2.69	0.9964	2.89	0.9981
2.10	**0.9821**	**2.30**	**0.9893**	**2.50**	**0.9938**	**2.70**	**0.9965**	**2.90**	**0.9981**
2.11	0.9826	2.31	0.9896	2.51	0.9940	2.71	0.9966	2.91	0.9982
2.12	0.9830	2.32	0.9894	2.52	0.9941	2.72	0.9967	2.92	0.9983
2.13	0.9834	2.33	0.9901	2.53	0.9943	2.73	0.9968	2.93	0.9983
2.14	0.9838	2.34	0.9904	2.54	0.9945	2.74	0.9969	2.94	0.9984
2.15	0.9842	2.35	0.9906	2.55	0.9946	2.75	0.9970	2.95	0.9984
2.16	0.9846	2.36	0.9909	2.56	0.9948	2.76	0.9971	2.96	0.9985
2.17	0.9850	2.37	0.9911	2.57	0.9949	2.77	0.9972	2.97	0.9985
2.18	0.9854	2.38	0.9913	2.58	0.9951	2.78	0.9973	2.98	0.9986
2.19	0.9857	2.39	0.9916	2.59	0.9952	2.79	0.9974	2.99	0.9986

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
3.00	**0.9987**	**3.20**	**0.9993**	**3.40**	**0.9997**	**3.60**	**0.9998**	**3.80**	**0.9999**
3.10	0.9990	3.30	0.9995	3.50	0.9998	3.70	0.9999	3.90	0.9999

21.18 χ^2 Distribution

For the formula of the χ^2 distribution see 16.2.4.6, p. 822.

χ^2 Distribution: Quantile $\chi^2_{\alpha,m}$

Degree of Freedom	**Probability α**					
m	0.99	0.975	0.95	0.05	0.025	0.01
1	0.00016	0.00098	0.0039	3.8	5.0	6.6
2	0.020	0.051	0.103	6.0	7.4	9.2
3	0.115	0.216	0.352	7.8	9.4	11.3
4	0.297	0.484	0.711	9.5	11.1	13.3
5	0.554	0.831	1.15	11.1	12.8	15.1
6	0.872	1.24	1.64	12.6	14.4	16.8
7	1.24	1.69	2.17	14.1	16.0	18.5
8	1.65	2.18	2.73	15.5	17.5	20.1
9	2.09	2.70	3.33	16.9	19.0	21.7
10	2.56	3.25	3.94	18.3	20.5	23.2
11	3.05	3.82	4.57	19.7	21.9	24.7
12	3.57	4.40	5.23	21.0	23.3	26.2
13	4.11	5.01	5.89	22.4	24.7	27.7
14	4.66	5.63	6.57	23.7	26.1	29.1
15	5.23	6.26	7.26	25.0	27.5	30.6
16	5.81	6.91	7.96	26.3	28.8	32.0
17	6.41	7.56	8.67	27.6	30.2	33.4
18	7.01	8.23	9.39	28.9	31.5	34.8
19	7.63	8.91	10.1	30.1	32.9	36.2
20	8.26	9.59	10.9	31.4	34.2	37.6
21	8.90	10.3	11.6	32.7	35.5	38.9
22	9.54	11.0	12.3	33.9	36.8	40.3
23	10.2	11.7	13.1	35.2	38.1	41.6
24	10.9	12.4	13.8	36.4	39.4	43.0
25	11.5	13.1	14.6	37.7	40.6	44.3
26	12.2	13.8	15.4	38.9	41.9	45.6
27	12.9	14.6	16.2	40.1	43.2	47.0
28	13.6	15.3	16.9	41.3	44.5	48.3
29	14.3	16.0	17.7	42.6	45.7	49.6
30	15.0	16.8	18.5	43.8	47.0	50.9
40	22.2	24.4	26.5	55.8	59.3	63.7
50	29.7	32.4	34.8	67.5	71.4	76.2
60	37.5	40.5	43.2	79.1	83.3	88.4
70	45.4	48.8	51.7	90.5	95.0	100.4
80	53.5	57.2	60.4	101.9	106.6	112.3
90	61.8	65.6	69.1	113.1	118.1	124.1
100	70.1	74.2	77.9	124.3	129.6	135.8

21.19 Fisher F Distribution

For the formula of the Fisher F distribution see 16.2.4.7, p. 823.

Fisher F Distribution: Quantile f_{α,m_1,m_2} for $\alpha = 0.05$

m_2	m_1											
	1	2	3	4	5	6	8	12	24	30	40	∞
1	161.4	199.5	215.7	224.6	230.2	234.0	238.9	243.9	249.0	250.0	251.0	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.37	19.41	19.45	19.46	19.47	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.85	8.74	8.64	8.62	8.59	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.04	5.91	5.77	5.75	5.72	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.82	4.68	4.53	4.50	4.46	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.15	4.00	3.84	3.81	3.77	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.73	3.57	3.41	3.38	3.34	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.44	3.28	3.12	3.08	3.05	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.23	3.07	2.90	2.86	2.83	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.07	2.91	2.74	2.70	2.66	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	2.95	2.79	2.61	2.57	2.53	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.85	2.69	2.51	2.47	2.43	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.77	2.60	2.42	2.38	2.34	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.70	2.53	2.35	2.31	2.27	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.64	2.48	2.29	2.25	2.20	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.59	2.42	2.24	2.19	2.15	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.55	2.38	2.19	2.15	2.10	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.51	2.34	2.15	2.11	2.06	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.48	2.31	2.11	2.07	2.03	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.45	2.28	2.08	2.04	1.99	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.42	2.25	2.05	2.01	1.96	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.40	2.23	2.03	1.98	1.94	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.37	2.20	2.00	1.96	1.91	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.36	2.18	1.98	1.94	1.89	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.34	2.16	1.96	1.92	1.87	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.32	2.15	1.95	1.90	1.85	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.31	2.13	1.93	1.88	1.84	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.29	2.12	1.91	1.87	1.82	1.65
29	4.18	3.33	2.93	2.70	2.55	2.43	2.28	2.10	1.90	1.85	1.80	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.27	2.09	1.89	1.84	1.79	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.18	2.00	1.79	1.74	1.69	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.10	1.92	1.70	1.65	1.59	1.39
125	3.92	3.07	2.68	2.44	2.29	2.17	2.01	1.83	1.60	1.55	1.49	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	1.94	1.75	1.52	1.46	1.39	1.00

Fisher F Distribution: Quantile f_{α,m_1,m_2} for $\alpha = 0,01$

m_2	m_1											
	1	2	3	4	5	6	8	12	24	30	40	∞
1	4052	4999	5403	5625	5764	5859	5981	6106	6235	6261	6287	6366
2	98.50	99.00	99.17	99.25	99.30	99.33	99.37	99.42	99.46	99.47	99.47	99.50
3	34.12	30.82	29.46	28.71	28.24	27.91	27.49	27.05	26.60	26.50	26.41	26.12
4	21.20	18.00	16.69	15.98	15.52	15.21	14.80	14.37	13.93	13.84	13.74	13.46
5	16.26	13.27	12.06	11.39	10.97	10.67	10.29	9.89	9.47	9.38	9.29	9.02
6	13.74	10.92	9.78	9.15	8.75	8.47	8.10	7.72	7.31	7.23	7.14	6.88
7	12.25	9.55	8.45	7.85	7.46	7.19	6.84	6.47	6.07	5.99	5.91	5.65
8	11.26	8.65	7.59	7.01	6.63	6.37	6.03	5.67	5.28	5.20	5.12	4.86
9	10.56	8.02	6.99	6.42	6.06	5.80	5.47	5.11	4.73	4.65	4.57	4.31
10	10.04	7.56	6.55	5.99	5.64	5.39	5.06	4.71	4.33	4.25	4.17	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.74	4.40	4.02	3.94	3.86	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.50	4.16	3.78	3.70	3.62	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.30	3.96	3.59	3.51	3.43	3.16
14	8.86	6.51	5.56	5.04	4.70	4.46	4.14	3.80	3.43	3.35	3.27	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.00	3.67	3.29	3.21	3.13	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	3.89	3.55	3.18	3.10	3.02	2.75
17	8.40	6.11	5.18	4.67	4.34	4.10	3.79	3.46	3.08	3.00	2.92	2.65
18	8.29	6.01	5.09	4.58	4.25	4.01	3.71	3.37	3.00	2.92	2.84	2.57
19	8.18	5.93	5.01	4.50	4.17	3.94	3.63	3.30	2.92	2.84	2.76	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.56	3.23	2.86	2.78	2.69	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.51	3.17	2.80	2.72	2.64	2.36
22	7.95	5.72	4.82	4.31	3.99	3.76	3.45	3.12	2.75	2.67	2.58	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.41	3.07	2.70	2.62	2.54	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.36	3.03	2.66	2.58	2.49	2.21
25	7.77	5.57	4.68	4.18	3.86	3.63	3.32	2.99	2.62	2.54	2.45	2.17
26	7.72	5.53	4.64	4.14	3.82	3.59	3.29	2.96	2.58	2.50	2.42	2.13
27	7.68	5.49	4.60	4.11	3.78	3.56	3.26	2.93	2.55	2.47	2.38	2.10
28	7.64	5.45	4.57	4.07	3.76	3.53	3.23	2.90	2.52	2.44	2.35	2.06
29	7.60	5.42	4.54	4.04	3.73	3.50	3.20	2.87	2.49	2.41	2.33	2.03
30	7.56	5.39	4.51	4.02	3.70	3.47	3.17	2.84	2.47	2.38	2.30	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	2.99	2.66	2.29	2.20	2.11	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.82	2.50	2.12	2.03	1.94	1.60
125	6.84	4.78	3.94	3.48	3.17	2.95	2.66	2.33	1.94	1.85	1.75	1.37
∞	6.63	4.60	3.78	3.32	3.02	2.80	2.51	2.18	1.79	1.70	1.59	1.00

21.20 Student t Distribution

For the formula of the Student t distribution see 16.2.4.8, p. 824.

Student t Distribution: Quantile $t_{\alpha,m}$ or $t_{\alpha/2,m}$

Degree of Freedom m	**Probability α for Two–Sided Problem**					
	0.10	0.05	0.02	0.01	0.002	0.001
1	6.31	12.7	31.82	63.7	318.3	637.0
2	2.92	4.30	6.97	9.92	22.33	31.6
3	2.35	3.18	4.54	5.84	10.22	12.9
4	2.13	2.78	3.75	4.60	7.17	8.61
5	2.01	2.57	3.37	4.03	5.89	6.86
6	1.94	2.45	3.14	3.71	5.21	5.96
7	1.89	2.36	3.00	3.50	4.79	5.40
8	1.86	2.31	2.90	3.36	4.50	5.04
9	1.83	2.26	2.82	3.25	4.30	4.78
10	1.81	2.23	2.76	3.17	4.14	4.59
11	1.80	2.20	2.72	3.11	4.03	4.44
12	1.78	2.18	2.68	3.05	3.93	4.32
13	1.77	2.16	2.65	3.01	3.85	4.22
14	1.76	2.14	2.62	2.98	3.79	4.14
15	1.75	2.13	2.60	2.95	3.73	4.07
16	1.75	2.12	2.58	2.92	3.69	4.01
17	1.74	2.11	2.57	2.90	3.65	3.96
18	1.73	2.10	2.55	2.88	3.61	3.92
19	1.73	2.09	2.54	2.86	3.58	3.88
20	1.73	2.09	2.53	2.85	3.55	3.85
21	1.72	2.08	2.52	2.83	3.53	3.82
22	1.72	2.07	2.51	2.82	3.51	3.79
23	1.71	2.07	2.50	2.81	3.49	3.77
24	1.71	2.06	2.49	2.80	3.47	3.74
25	1.71	2.06	2.49	2.79	3.45	3.72
26	1.71	2.06	2.48	2.78	3.44	3.71
27	1.71	2.05	2.47	2.77	3.42	3.69
28	1.70	2.05	2.46	2.76	3.40	3.66
29	1.70	2.05	2.46	2.76	3.40	3.66
30	1.70	2.04	2.46	2.75	3.39	3.65
40	1.68	2.02	2.42	2.70	3.31	3.55
60	1.67	2.00	2.39	2.66	3.23	3.46
120	1.66	1.98	2.36	2.62	3.17	3.37
∞	1.64	1.96	2.33	2.58	3.09	3.29
	0.05	0.025	0.01	0.005	0.001	0.0005
	Probability α for One–Sided Problem					

21.21 Random Numbers

For the meaning of random numbers see 16.3.5.2, p. 843.

4730	1530	8004	7993	3141	0103	4528	7988	4635	8478	9094	9077	5306	4357	8353
0612	2278	8634	2549	3737	7686	0723	4505	6841	1379	6460	1869	5700	5339	6862
0285	1888	9284	3672	7033	4844	0149	7412	6370	1884	0717	5740	8477	6583	0717
7768	9078	3428	2217	0293	3978	5933	1032	5192	1732	2137	9357	5941	6564	2171
4450	8085	8931	3162	9968	6369	1256	0416	4326	7840	6525	2608	5255	4811	3763
7332	6563	4013	7406	4439	5683	6877	2920	9588	3002	2869	3746	3690	6931	1230
4044	1643	9005	5969	9442	7696	7510	1620	4973	1911	1288	6160	9797	8755	6120
0067	7697	9278	4765	9647	4364	1037	4975	1998	1359	1346	6125	5078	6742	3443
5358	5256	7574	3219	2532	7577	2815	8696	9248	9410	9282	6572	3940	6655	9014
0038	4772	0449	6906	8859	5044	8826	6218	3206	9034	0843	9832	2703	8514	4124
8344	2271	4689	3835	2938	2671	4691	0559	8382	2825	4928	5379	8635	8135	7299
7164	7492	5157	8731	4980	8674	4506	7262	8127	2022	2178	7463	4842	4414	0127
7454	7616	8021	2995	7868	0683	3768	0625	9887	7060	0514	0034	8600	3727	5056
3454	6292	0067	5579	9028	5660	5006	8325	9677	2169	3196	0357	7811	5434	0314
0401	7414	3186	3081	5876	8150	1360	1868	9265	3277	8465	7502	6458	7195	9869
6202	0195	1077	7406	4439	5683	6877	2920	9588	3002	2869	3746	3690	2705	6251
8284	0338	4286	5969	9442	7696	7510	1620	6973	1911	1288	6160	9797	1547	4972
9056	0151	7260	4765	9647	4364	1037	4975	1998	1359	1346	6125	5078	3424	1354
9747	3840	7921	3219	2532	7577	2815	8696	9248	9410	9282	6572	3940	8969	3659
2992	8836	3342	6906	8859	5044	8826	6218	3206	9034	0843	9832	2703	5225	8898
6170	4595	2539	7592	1339	4802	5751	3785	7125	4922	8877	9530	6499	6432	1516
3265	8619	0814	5133	7995	8030	7408	2186	0725	5554	5664	6791	9677	3085	8319
0179	3949	6995	3170	9915	6960	2621	6718	4059	9919	1007	6469	5410	0246	3687
1839	6042	9650	3024	0680	1127	8088	0200	5868	0084	6362	6808	3727	8710	6065
2276	8078	9973	4398	3121	7749	8191	2087	8270	5233	3980	6774	8522	5736	3132
4146	9952	7945	5207	1967	7325	7584	3485	5832	8118	8433	0606	2719	2889	2765
3526	3809	5523	0648	3326	1933	6265	0649	6177	2139	7236	0441	1352	1499	3068
3390	7825	7012	9934	7022	2260	0190	1816	7933	2906	3030	6032	1685	3100	1929
4806	9286	5051	4651	1580	5004	8981	1950	2201	3852	6855	5489	6386	3736	0498
7959	5983	0204	4325	5039	7342	7252	2800	4706	6881	8828	2785	8375	7232	2483
8245	9611	0641	7024	3899	8981	1280	5678	8096	7010	1435	7631	7361	8903	8684
7551	4915	2913	9031	9735	7820	2478	9200	7269	6284	9861	2849	2208	8616	5865
5903	2744	7318	7614	5999	1246	9759	6565	1012	0059	2419	0036	2027	5467	5577
9001	4521	5070	4150	5059	5178	7130	2641	7812	1381	6158	9539	3356	5861	9371
0265	3305	3814	0973	4958	4830	6297	0575	4843	3437	5629	3496	5406	4790	9734

22 Bibliography

1. Arithmetic

[1.1] Beckenbach, E.; Bellmann, R.: Inequalities. — Springer-Verlag 1983.
[1.2] Bosch, K.: Finanzmathematik. — Oldenbourg-Verlag 1991.
[1.3] Hardy, G.: A Course in Pure Mathematics. — Cambridge University Press 1952.
[1.4] Heilmann, W.–R.: Grundbegriffe der Risikotheorie. — Verlag Versicherungswirtschaft 1986.
[1.5] Isenbart, F.; Münzer, H.: Lebensversicherungsmathematik für Praxis und Studium. — Verlag Gabler, 2nd ed. 1986.
[1.6] Gellert, W.; Kästner, H.; Neuber, S.: Fachlexikon ABC Mathematik. — Verlag H. Deutsch 1978.
[1.7] Heitzinger, W.; Troch, I.; Valentin, G.: Praxis nichtlinearer Gleichungen. — C. Hanser Verlag 1984.
[1.8] Pfeifer, A.: Praktische Finanzmathematik. — Verlag H. Deutsch 1995.

2. Functions

[2.1] Fetzer, A.; Fränkel, H.: Mathematik Lehrbuch für Fachhochschulen, Bd. 1. — VDI-Verlag 1995.
[2.2] Fichtenholz, G.M.: Differential- und Integralrechnung, Bd. 1. — Verlag H. Deutsch 1994.
[2.3] Hardy, G.: A Course in Pure Mathematics. — Cambridge University Press 1952.
[2.4] Handbook of Mathematical, Scientific and Engineering Formulas, Tables, Functions, Graphs, Transforms. — Research and Education Association 1961.
[2.5] Papula, L.: Mathematik für Ingenieure, Bd. 1, 2, 3. — Verlag Vieweg 1994–1996.
[2.6] Smirnow, W.I.: Lehrbuch der höheren Mathematik, Bd. 1. — Verlag H. Deutsch 1994.

3. Geometry

[3.1] Bär, G.: Geometrie. — B. G. Teubner 1996.
[3.2] Berger, M.: Geometry, Vol. 1, 2. — Springer-Verlag 1987.
[3.3] Böhm, J.: Geometrie, Bd. 1, 2. — Verlag H. Deutsch 1988.
[3.4] Dreszer, J.: Mathematik Handbuch für Technik und Naturwissenschaft. — Verlag H. Deutsch 1975.
[3.5] Dubrovin, B.; Fomenko, A.; Novikov, S.: Modern Geometry, Vol. 1–3. — Springer-Verlag 1995.
[3.6] Efimow, N.V.: Höhere Geometrie, Bd. 1, 2. — Verlag Vieweg 1970.
[3.7] Fischer, G.: Analytische Geometrie. — Verlag Vieweg 1988.
[3.8] Jennings, G.A.: Modern Geometry with Applications. — Springer-Verlag 1994.
[3.9] Lang, S.; Murrow, G.: A High School Course. — Springer-Verlag 1991.
[3.10] Gellert, W.; Küstner, H.; Kästner, H. (Eds.): Kleine Enzyklopädie Mathematik. — Verlag H. Deutsch 1988.
[3.11] Klingenberg, W.: Lineare Algebra und Geometrie. — Springer-Verlag 1992.
[3.12] Klotzek, B.: Einführung in die Differentialgeometrie, Bd. 1, 2. — Verlag H. Deutsch 1995.
[3.13] Koecher, M.: Lineare Algebra und analytische Geometrie. — Springer-Verlag 1997.
[3.14] Mangoldt, H. v.; Knopp, K.: Einführung in die höhere Mathematik, Bd. II. — S. Hirzel Verlag 1978.
[3.15] Matthews, V.: Vermessungskunde Teil 1, 2. — B. G. Teubner 1993.
[3.16] Raschewski, P.K.: Riemannsche Geometrie und Tensoranalysis. — Verlag H. Deutsch 1995.
[3.17] Sigl, R.: Ebene und sphärische Trigonometrie. — Verlag H. Wichmann 1977.
[3.18] Singer, D. A.: Plane and Fancy. — Springer-Verlag 1998.

I.N. Bronshtein et al., *Handbook of Mathematics*,
DOI 10.1007/978-3-662-46221-8_22

[3.19] Steinert, K.-G.: Sphärische Trigonometrie. — B. G. Teubner 1977.
[3.22] Zhigang Xiang, Plastock, R.A.: Computergraphik. — mitp–Verlag Bonn, 2003.

4. Linear Algebra

[4.1] Berendt, G.; Weimar, E.: Mathematik für Physiker, Bd. 1, 2. — VCH 1990.
[4.2] Blyth, T. S.; Robertson, E. F.: Basic Linear Algebra. — Springer-Verlag 1998.
[4.3] Curtis, C. W.: Linear Algebra. An Introductory Approach. — Springer-Verlag 1984.
[4.4] Faddejew, D.K.; Faddejewa, W.N.: Numerische Methoden der linearen Algebra. — Deutscher Verlag der Wissenschaften 1970.
[4.5] Gürlebeck, K.; Habetha, K.; Sprössig, W.: Funktionentheorie in der Ebene und im Raum. — Birkhäuser–Verlag 2006.
[4.6] Jänich, K.: Lineare Algebra. — Springer-Verlag 1996.
[4.7] Kiełbasiński, A.; Schwetlick, H.: Numerische lineare Algebra. Eine computerorientierte Einführung. — Verlag H. Deutsch 1988.
[4.8] Klingenberg, W.: Lineare Algebra und Geometrie. — Springer-Verlag 1992.
[4.9] Koecher, M.: Lineare Algebra und analytische Geometrie. — Springer-Verlag 1997.
[4.10] Lippmann, H.: Angewandte Tensorrechnung. Für Ingenieure, Physiker und Mathematiker. — Springer-Verlag 1993.
[4.11] Raschewski, P.K.: Riemannsche Geometrie und Tensoranalysis. — Verlag H. Deutsch 1995.
[4.12] Schneider, P.J.;Eberly, D.H.: Geometric Tools for Computer Graphics. Morgan Kaufmann, San Francisco 2003.
[4.13] Smirnow, W.I.: Lehrbuch der höheren Mathematik, Teil III,1. — Verlag H. Deutsch 1994.
[4.14] Smith, L.: Lineare Algebra. — Springer-Verlag 1998.
[4.15] Zurmühl, R.; Falk, S.: Matrizen und ihre Anwendung – 1. Grundlagen. — Springer-Verlag 1997.
[4.16] Zurmühl, R.: Praktische Mathematik für Ingenieure und Physiker. — Springer-Verlag 1984.

5. Algebra and Discrete Mathematics

A) Algebra and Discrete Mathematics, General

[5.1] Aigner, M.: Diskrete Mathematik. — Verlag Vieweg 1993.
[5.2] Burris, S.; Sankappanavar, H. P.: A Course in Universal Algebra. — Springer-Verlag 1981.
[5.3] Ehrig, H.; Mahr, B.: Fundamentals of Algebraic Specification 1. — Springer-Verlag 1985.
[5.4] Gries, D.; Schneider, F.: A Logical Approach to Discrete Mathematics. — Springer-Verlag 1993.
[5.5] Wechler, W.: Universal Algebra for Computer Scientists. — Springer-Verlag 1992.

B) Algebra and Discrete Mathematics, Group Theory

[5.6] Fässler, A.; Stiefel, E.: Group Theoretical Methods and their Applications. — Birkhäuser 1992.
[5.7] Hein, W.: Struktur und Darstellungstheorie der klassischen Gruppen. — Springer-Verlag 1990.
[5.8] Heine, V.: Group Theory in Quantum Mechanics. — Dover 1993.
[5.9] Lucha, W.; Schöberl, F.F.: Gruppentheorie. — B.I. Wissenschaftsverlag 1993.
[5.10] Ludwig, W., Falter, C.: Symmetries in Physics. Group Theory Applied to Physical Problems. — Springer-Verlag 1996.
[5.11] Varadarajan, V.: Lie Groups, Lie Algebras and their Representation. — Springer-Verlag 1990.
[5.12] Wallace, D.: Groups, Rings and Fields. — Springer-Verlag 1998.

[5.13] Zachmann, H.G.: Mathematik für Chemiker. — VCH 1990.

C) Algebra and Discrete Mathematics, Number Theory

[5.14] Jones, G.A.; Jones, J.M.: Elementary Number Theory. — Springer-Verlag.
[5.15] Nathanson, M.: Elementary Methods in Number Theory. — Springer-Verlag.
[5.16] Rivest, R.L.; Shamir, A.; Adleman, L.: A Method for Obtaining Digital Signatures and Public Key Cryptosystems. — Comm. ACM **21** (1978) 12–126.

D) Algebra and Discrete Mathematics, Cryptology

[5.17] Bauer, F. L.: Decrypted Secrets. — Springer-Verlag.
[5.18] Beutelspacher, A.: Cryptology. The Mathematical Association of America 1996.
[5.19] Schneider, B.: Applied Cryptology. — John Wiley 1995.
[5.20] Stallings, W.: Cryptology and Network Security. Prentice Hall 1998. — Addison Wesley Longman 1997.
[5.21] http://csrc.nist.gov/publications/fips/fips46-3/fips46-3.pdf
[5.22] http://csrc.nist.gov/publications/fips/fips197/fips197.pdf

E) Algebra and Discrete Mathematics, Graph Theory

[5.23] Diestel, R.: Graph Theory. — Springer-Verlag.
[5.24] Edmonds, J.: Paths, Trees and Flowers. — Canad. J. Math. **17** (1965) 449–467.
[5.25] Edmonds, J., Johnson, E.L.: Matching, Euler Tours and the Chinese Postman. — Math. Programming **5** (1973) 88–129.
[5.26] Harary, F.: Graph Theory. — Addison Wesley.

F) Algebra and Discrete Mathematics, Fuzzy–Logik

[5.27] Bandemer, H.; Gottwald, S.: Einführung in Fuzzy-Methoden – Theorie und Anwendungen unscharfer Mengen. — Akademie-Verlag 1993.
[5.28] Driankov, D.; Hellendorn, H.; Reinfrank, M.: An Introduction to Fuzzy Control. — Springer-Verlag 1993.
[5.29] Dubois, D.; Prade, H.: Fuzzy Sets and System Theory and Applications. — Academic Press 1980.
[5.30] Grauel, A.: Fuzzy-Logik. Einführung in die Grundlagen mit Anwendungen. — B.I. Wissenschaftsverlag 1995.
[5.31] Hordeson, J.N.; Nair, N.S.: Fuzzy Mathematics. – An Introduction for Engineers and Scientists. — Physica Verlag 1998.
[5.32] Kruse, R.; Gebhardt, J.; Klawonn, F.: Fuzzy-Systeme. — B. G. Teubner 1993.
[5.33] Wang, Z.; Klir, G.T.: Fuzzy Measure Theory. — Plenum Press 1992.
[5.34] Zimmermann, H-J.: Fuzzy Sets. Decision Making and Expert Systems. — Verlag Kluwer–Nijhoff 1987.

6. Differential Calculus

[6.1] Courant, R.: Introduction to Calculus and Analysis, Vols. 1 and 2. — Springer-Verlag 1989.
[6.2] Fetzer, A.; Fränkel, H.: Mathematik. – Lehrbuch für Fachhochschulen, Bd. 1, 2. — VDI–Verlag 1995.
[6.3] Fichtenholz, G.M.: Differential- und Integralrechnung, Bd. 1–3. — Verlag H. Deutsch 1994.
[6.4] Lang, S.: Calculus of Several Variables. — Springer-Verlag 1987.
[6.5] Knopp, K.: Theorie und Anwendung der unendlichen Reihen. — Springer-Verlag 1964.
[6.6] Mangoldt, H. v.; Knopp, K.: Einführung in die höhere Mathematik, Bd. 2, 3. — S. Hirzel Verlag 1978–81.
[6.7] Papula, L.: Mathematik für Ingenieure, Bd. 1–3. — Verlag Vieweg 1994–1996.
[6.8] Smirnow, W.I.: Lehrbuch der höheren Mathematik, Bd. II, III. — Verlag H. Deutsch 1994.

[6.9] ZACHMANN, H.G.: Mathematik für Chemiker. — VCH 1990.

7. Infinite Series

[7.1] APELBLAT, A.: Tables of Integrals and Series. — Verlag H. Deutsch 1996.
[7.2] COURANT, R.: Introduction to Calculus and Analysis, Vols. 1 and 2. — Springer-Verlag 1989.
[7.3] FETZER, A.; FRÄNKEL, H.: Mathematik. – Lehrbuch für Fachhochschulen, Bd. 1, 2. — VDI-Verlag 1995.
[7.4] FICHTENHOLZ, G.M.: Differential- und Integralrechnung, Bd. 1–3. — Verlag H. Deutsch 1994.
[7.5] KNOPP, K.: Theorie und Anwendung der unendlichen Reihen. — Springer-Verlag 1964.
[7.6] MANGOLDT, H. V.; KNOPP, K.; HRG. F. LÖSCH: Einführung in die höhere Mathematik, Bd. 1–4. — S. Hirzel Verlag 1989.
[7.7] PAPULA, L.: Mathematik für Ingenieure, Bd. 1–3. — Verlag Vieweg 1994–1996.
[7.8] PLASCHKO, P.; BROD, K.: Höhere mathematische Methoden für Ingenieure und Physiker. —Springer-Verlag 1989.
[7.9] SMIRNOW, W.I.: Lehrbuch der höheren Mathematik, Bd. II, III. — Verlag H. Deutsch 1994.

8. Integral Calculus

[8.1] APELBLAT, A.: Tables of Integrals and Series. — Verlag H. Deutsch 1996.
[8.2] BRYTSCHKOW, J.A.; MARITSCHEW, O.I.; PRUDNIKOV, A.P.: Tabellen unbestimmter Integrale. — Verlag H. Deutsch 1992.
[8.3] COURANT, R.: Introduction to Calculus and Analysis, Vols. 1 and 2. — Springer-Verlag 1989.
[8.4] FICHTENHOLZ, G.M.: Differential- und Integralrechnung, Bd. 1–3. — Verlag H. Deutsch 1994.
[8.5] KAMKE, E.: Das Lebesgue–Stieltjes-Integral. — B. G. Teubner 1960.
[8.6] KNOPP, K.: Theorie und Anwendung der unendlichen Reihen. — Springer-Verlag 1964.
[8.7] MANGOLDT, H. V.; KNOPP, K.; HRSG. F. LÖSCH: Einführung in die höhere Mathematik, Bd. 1–4. — S. Hirzel Verlag 1989.
[8.8] PAPULA, L.: Mathematik für Ingenieure, Bd. 1–3. — Verlag Vieweg 1994–1996.
[8.9] SMIRNOW, W.I.: Lehrbuch der höheren Mathematik, Bd. II, III. — Verlag H. Deutsch 1994.
[8.10] ZACHMANN, H.G.: Mathematik für Chemiker. — VCH 1990.

9. Differential Equations

A) Ordinary and Partial Differential Equations

[9.1] BRAUN, M.: Differentialgleichungen und ihre Anwendungen. — Springer-Verlag 1991.
[9.2] CODDINGTON, E.; LEVINSON, N.: Theory of Ordinary Differential Equations. — McGraw Hill 1955.
[9.3] COLLATZ, L.: Differentialgleichungen. — B. G. Teubner 1990.
[9.4] COLLATZ, L.: Eigenwertaufgaben mit technischen Anwendungen. — Akademische Verlagsgesellschaft 1963.
[9.5] COURANT, R.; HILBERT, D.: Methoden der mathematischen Physik, Bd. 1, 2. — Springer-Verlag 1968.
[9.6] EGOROV, YU.; SHUBIN, M.: Partial Differential Equations, Vols. 1–4. – Encyclopaedia of Mathematical Sciences. Springer-Verlag 1991.
[9.7] FRANK, PH.; MISES, R. V.: Die Differential- und Integralgleichungen der Mechanik und Physik, Bd. 1, 2. — Verlag Vieweg 1961.
[9.8] GREINER, W.: Quanten Mechanics. An Introduction. — Springer-Verlag 1994.
[9.9] KAMKE, E.: Differentialgleichungen, Lösungsmethoden und Lösungen, Teil 1, 2. — BSB B. G. Teubner 1977.
[9.10] LANDAU, L.D.; LIFSCHITZ, E.M.: Quantenmechanik. — Verlag H. Deutsch 1992.

[9.11] Poljanin, A.D.; Saizew, V.F.: Sammlung gewöhnlicher Differentialgleichungen. — Verlag H. Deutsch 1996.
[9.12] Reissig, R.; Sansone, G.; Conti, R.: Nichtlineare Differentialgleichungen höherer Ordnung. — Edizioni Cremonese 1969.
[9.13] Smirnow, W.I.: Lehrbuch der höheren Mathematik, Teil 2. — Verlag H. Deutsch 1994.
[9.14] Sommerfeld, A.: Partielle Differentialgleichungen der Physik. — Verlag H. Deutsch 1992.
[9.15] Stepanow, W.W.: Lehrbuch der Differentialgleichungen. — Deutscher Verlag der Wissenschaften 1982.

B) Non-Linear Partial Differential Equations

[9.16] Akhmediev, N.N.;Ankiewicz, A. (Eds.): Dissipative Solitons. Lect. Notes Phys. — Springer–Verlag, Berlin 2005.
[9.17] Carr, L.D.; Reinhardt, W.P.: Phys. Rev. A **62**, 063610 (2000), *ibid.* **62**, 063611 (2000).
[9.18] Dodd, R.K.; Eilbeck, J.C.; Gibbon, J.D.; Morris, H.C.: Solitons and Non-Linear Wave Equations. — Academic Press 1982.
[9.19] Drazin, P.G., Johnson, R.: Solitons. An Introduction. — Cambridge University Press 1989.
[9.20] Gu Chaohao (Ed.): Soliton Theory and its Applications. — Springer-Verlag 1995.
[9.21] Lamb, G.L.: Elements of Soliton Theory. — John Wiley 1980.
[9.22] Makhankov, V.G.: Soliton Phenomenology. — Verlag Kluwer 1991.
[9.23] Pitaevskii, L.; Stringari, S.: Bose–Einstein condensation. — Oxford University Press 2003.
[9.24] Remoissenet, S.: Waves Called Solitons. Concepts and Experiments. — Springer-Verlag 1994.
[9.25] Toda, M.: Nonlinear Waves and Solitons. — Verlag Kluwer 1989.
[9.26] Tsuzuki, T.: Low Temp. Phys. **4**, 441 (1971).
[9.27] Vvedensky, D.: Partical Differential Equations with Mathematica. — Addison Wesley 1993.
[9.28] Wladimirow, V.S.: Gleichungen der mathematischen Physik. — Deutscher Verlag der Wissenschaften 1972.
[9.29] Ziesche, P.; Lehmann, G.: Elektronentheorie der Metalle. - Springer, Berlin 1983, S. 532-543.
[9.30] Ziesche, P.: Proof of an Addition Theorem for the Spherical von Neumann Functions Using Kasterin's Formula. - ZAMM **52**, 375 (1972).
[9.31] Ziesche, P.: Certain Sum Rules for Spherical Bessel Functions. - ZAMM, **57**, 194 (1977).

10. Calculus of Variations

[10.1] Blanchard, P.; Brüning, E.: Variational Methods in Mathematical Physics. — Springer-Verlag 1992.
[10.2] Giaquinta, M.; Hildebrandt, S.: Calculus of Variations. — Springer-Verlag 1995.
[10.3] Klingbeil, E.: Variationsrechnung. — BI-Verlag 1988.
[10.4] Klötzler, R.: Mehrdimensionale Variationsrechnung. — Birkhäuser 1970.
[10.5] Kosmol, P.: Optimierung und Approximation. — Verlag W. de Gruyter 1991.
[10.6] Michlin, S.G.: Numerische Realisierung von Variationsmethoden. — Akademie-Verlag 1969.
[10.7] Rothe, R.: Höhere Mathematik für Mathematiker, Physiker, Ingenieure, Teil VII. — B. G. Teubner 1960.

11. Linear Integral Equations

[11.1] Corduneanu, I.C.: Integral Equations and Applications. — Cambridge University Press 1991.
[11.2] Estrada, R.; Kanwal, R.P.: Singular Integral Equations. — John Wiley 1999.
[11.3] Hackbusch, W.: Integral Equations: Theory and Numerical Treatment. — Springer-Verlag 1995.
[11.4] Kanwal, R.P.: Linear Integral Equations. — Springer-Verlag 1996.

[11.5] KRESS, R.: Linear Integral Equations. — Springer-Verlag 1999.
[11.6] MICHLIN, S.G.; PRÖSSDORF, S.: Singular Integral Operators. — Springer-Verlag 1986.
[11.7] MICHLIN, S.G.: Integral Equations and their Applications to Certain Problems in Mechanics. — MacMillan 1964.
[11.8] MUSKELISHVILI, N.I.: Singular Integral Equations: Boundary Problems of Functions Theory and their Applications to Mathematical Physics. — Dover 1992.
[11.9] PIPKIN, A.C.: A Course on Integral Equations. — Springer-Verlag 1991.
[11.10] POLYANIN, A.D.; MANZHIROV, A.V.: Handbook of Integral Equations. — CRC Press 1998.
[11.11] VON SCHMEIDLER, W.: Integralgleichungen mit Anwendungen in Physik und technik. — Akademische Verlagsgesellschaft 1950.
[11.12] SMIRNOW, W.I.: Lehrgang der hoheren Mathematik, Bd. IV/1 — Verlag Harri Deutsch 1994.

12. Functional Analysis

[12.1] ACHIESER, N.I.; GLASMANN, I.M.: Theory of Linear Operators in Hilbert Space. — M. Nestell. Ungar. 1961.
[12.2] ALIPRANTIS, C.D.; BURKINSHAW, O.: Positive Operators. — Academic Press 1985.
[12.3] ALIPRANTIS, C.D.; BORDER, K.C.; LUXEMBURG, W.A.J.: Positive Operators, Riesz Spaces and Economics. — Springer-Verlag 1991.
[12.4] ALT, H.W.: Lineare Funktionalanalysis. – Eine anwendungsorientierte Einführung. — Springer-Verlag 1976.
[12.5] BALAKRISHNAN, A.V.: Applied Functional Analysis. — Springer-Verlag 1976.
[12.6] BAUER, H.: Maß- und Integrationstheorie. — Verlag W. de Gruyter 1990.
[12.7] BRONSTEIN, I.N.; SEMENDAJEW, K.A.: Ergänzende Kapitel zum Taschenbuch der Mathematik. — BSB B. G. Teubner 1970; Verlag H. Deutsch 1990.
[12.8] DUNFORD, N.; SCHWARTZ, J.T.: Linear Operators, Vols. I, II, III. — Intersciences 1958, 1963, 1971.
[12.9] EDWARDS, R.E.: Functional Analysis. — Holt, Rinehart and Winston 1965.
[12.10] GAJEWSKI, H.; GRÖGER, K.; ZACHARIAS, K.: Nichtlineare Operatorengleichungen und Operatordifferentialgleichungen. — Akademie-Verlag 1974.
[12.11] HALMOS, P.R.: A Hilbert Space Problem Book. — Van Nostrand 1967.
[12.12] HUTSON, V.C.L.; PYM, J.S.: Applications of Functional Analysis and Operator Theory. — Academic Press 1980.
[12.13] HEWITT, E.; STROMBERG, K.: Real and Abstract Analysis. — Springer-Verlag 1965.
[12.14] JOSHI, M.C.; BOSE, R.K.: Some Topics in Nonlinear Functional Analysis. — Wiley Eastern 1985.
[12.15] KANTOROVICH, L.V.; AKILOW, G.P.: Functional Analysis — Pergamon Press 1982.
[12.16] KOLMOGOROW, A.N.; FOMIN, S.W.: Introduction to Functional Analysis. — Graylock Press 1961.
[12.17] KRASNOSEL'SKIJ, M.A.; LIFSHITZ, J.A., SOBOLEV, A.V.: Positive Linear Systems. — Heldermann-Verlag 1989.
[12.18] LUSTERNIK, L.A.; SOBOLEW, V.I.: Elements of Functional Analysis. — Gordon and Breach 1961, Hindustan Publishing Corporation Delhi 1974, in German: Verlag H. Deutsch 1975.
[12.19] MEYER-NIEBERG, P.: Banach Lattices. — Springer-Verlag 1991.
[12.20] NAIMARK, M.A.: Normed Rings. — Wolters–Noordhoff 1972.
[12.21] RUDIN, W.: Functional Analysis. — McGraw–Hill 1973.
[12.22] SCHAEFER, H.H.: Topological Vector Spaces. — Macmillan 1966.
[12.23] SCHAEFER, H.H.: Banach Lattices and Positive Operators. — Springer-Verlag 1974.
[12.24] YOSIDA, K.: Functional Analysis. — Springer-Verlag 1965.

13. Vector Analysis and Vector Fields

[13.1] Domke, E.: Vektoranalysis: Einführung für Ingenieure und Naturwissenschaftler. — BI-Verlag 1990.

[13.2] Jänich, K.: Vector Analysis. — Springer-Verlag 1999.

[13.3] Schark, R.: Vektoranalysis für Ingenieurstudenten. — Verlag H. Deutsch 1992.

14. Function Theory

[14.1] Abramowitz, M.; Stegun, I. A.: Pocketbook of Mathematical Functions. — Verlag H. Deutsch 1984.

[14.2] Behnke, H.; Sommer, F.: Theorie der analytischen Funktionen einer komplexen Veränderlichen. — Springer-Verlag 1976.

[14.3] Fichtenholz, G.M.: Differential- und Integralrechnung, Bd. 2. — Verlag H. Deutsch 1994.

[14.4] Freitag, E.; Busam, R.: Funktionentheorie. — Springer-Verlag 1994.

[14.5] Greuel, O.; Kadner, H.: Komplexe Funktionen und konforme Abbildungen. — B. G. Teubner 1990.

[14.6] Jahnke, E.; Emde, F.: Tafeln höherer Funktionen. — B. G. Teubner 1960.

[14.7] Jänich, K.: Funktionentheorie. Eine Einführung. — Springer-Verlag 1993.

[14.8] Knopp, Soni, R.: Funktionentheorie. — Verlag W. de Gruyter 1976.

[14.9] Magnus, W.; Oberhettinger, F.: Formulas and Theorems for the Special Functions of mathematical Physics, 3rd ed. — Springer-Verlag 1966.

[14.10] Oberhettinger, F.; Magnus, W.: Anwendung der elliptischen Funktionen in Physik und Technik. — Springer-Verlag 1949.

[14.11] Schark, R.: Funktionentheorie für Ingenieurstudenten. — Verlag H. Deutsch 1993.

[14.12] Smirnow: Lehrbuch der höheren Mathematik, Bd. III. — Verlag H. Deutsch 1994.

[14.13] Springer, S.: Introduction to Riemann Surfaces. — Chelsea Publishing Company 1981.

15. Integral Transformations

[15.1] Blatter, C.: Wavelets. – Eine Einführung. — Vieweg 1998.

[15.2] Doetsch, G.: Introduction to the Theory and Application of the Laplace Transformation. — Springer-Verlag 1974.

[15.3] Dyke, P.P.G.: An Introduction to Laplace Transforms and Fourier Series. — Springer-Verlag 2000.

[15.4] Fetzer, V.: Integral-Transformationen. — Hüthig 1977.

[15.5] Föllinger, O.: Laplace- und Fourier-Transformation. — Hüthig 1993.

[15.6] Gauss, E.: WALSH-Funktionen für Ingenieure und Naturwissenschaftler. — B. G. Teubner 1994.

[15.7] Hubbard, B.B.: Wavelets. Die Mathematik der kleinen Wellen. — Birkhäuser 1997.

[15.8] Jennison, R.C.: Fourier Transforms and Convolutions for the Experimentalist. — Pergamon Press 1961.

[15.9] Louis, A. K.; Maass, P.; Rieder, A.: Wavelets. Theorie und Anwendungen. — B. G. Teubner 1994.

[15.10] Oberhettinger, F.: Tables of Fourier Transforms of Distributions. — Springer-Verlag 1990.

[15.11] Oberhettinger, F.; Badil, L.: Tables of Laplace Transforms. — Springer-Verlag 1973.

[15.12] Papoulis, A.: The Fourier Integral and its Applications. — McGraw Hill 1962.

[15.13] Schiff, J.L.: The Laplace Transform. – Theory and Applications. — Springer-Verlag 1999.

[15.14] Sirovich, L.: Introduction to Applied Mathematics. — Springer-Verlag 1988.

[15.15] Tolimieri, R.; An, M.; Lu, C.: Mathematics of Multidimensional Fourier Transforms Algorithms. — Springer-Verlag 1997.

[15.16] Tolimieri, R.; An, M.; Lu, C.: Algorithms for Discrete Transform and Convolution. — Springer-Verlag 1997.

[15.17] Voelker, D.; Doetsch, G.: Die zweidimensionale Laplace-Transformation. — Birkhäuser 1950.

[15.18] Walker, J.S.: Fast Fourier Transforms. — Springer-Verlag 1996.

16. Probability Theory and Mathematical Statistics

[16.1] Berger, M.,A.: An Introduction to Probability and Stochastic Processes. — Springer-Verlag 1993.
[16.2] Behnen, K.; Neuhaus, G.: Grundkurs Stochastik. — B. G. Teubner 1995.
[16.3] Brandt, S.: Data Analysis. Statistical and Computational Methods for Scientists and Engineers. — Springer-Verlag 1999.
[16.4] Clauss, G.; Finze, F.-R.; Partzsch, L.: Statistik für Soziologen, Pädagogen, Psychologen und Mediziner, Bd. 1. — Verlag H. Deutsch 1995.
[16.5] Fisz, M.: Wahrscheinlichkeitsrechnung und mathematische Statistik. — Deutscher Verlag der Wissenschaften 1988.
[16.6] Gardiner, C.W.: Handbook of Stochastic Methods. — Springer-Verlag 1997.
[16.7] Gnedenko, B.W.: Lehrbuch der Wahrscheinlichkeitstheorie. — Verlag H. Deutsch 1997.
[16.8] Hartmann; Lezki; Schäfer: Mathematische Methoden in der Stoffwirtschaft. — Deutscher Verlag für Grundstoffindustrie.
[16.9] Hübner, G.: Stochastik. – Eine anwendungsorientierte Einführung für Informatiker, Ingenieure und Mathematiker. — Vieweg, 3rd ed. 2000.
[16.10] Koch, K.-R.: Parameter Estimation and Hypothesis Testing in Linear Models. — Springer-Verlag 1988.
[16.11] Kolmogoroff, A.N.: Grundbegriffe der Wahrscheinlichkeitsrechnung. — Springer–Verlag 1933, 1977.
[16.12] Rinne, H.: Taschenbuch der Statistik. — Verlag H. Deutsch 1997.
[16.13] Shao, Jun: Mathematical Statistics. — Springer-Verlag 1999.
[16.14] Siniai, J.G.: Probability Theory. — Springer-Verlag 1992.
[16.15] Sobol, I.M.: Die Monte-Carlo-Methode. — Verlag H. Deutsch 1991.
[16.16] Storm, R.: Wahrscheinlichkeitsrechnung, mathematische Statistik und statistische Qualitätskontrolle. — Fachbuchverlag 1995.
[16.17] Taylor, J.R.: An Introduction to Error Analysis. — University Science Books 1982, VCH 1988.
[16.18] Terrell, G.R.: Mathematical Statistics. A Unified Introduction. — Springer-Verlag 1999.
[16.19] Weber, H.: Einführung in die Wahrscheinlichkeitsrechnung und Statistik für Ingenieure. — B. G. Teubner 1992.
[16.20] Weber, E.: Grundriß der biologischen Statistik für Naturwissenschaftler, Landwirte und Mediziner. — Gustav Fischer Verlag 1972.

17. Dynamical Systems and Chaos

[17.1] Arrowsmith, D.K.; Place, C.M.: An introduction to Dynamical Systems. — Cambridge University Press 1990.
[17.2] Falconer, K.: Fractal Geometry. — John Wiley 1990.
[17.3] Guckenheimer, J.; Holmes, P.: Non-Linear Oscillations, Dynamical Systems and Bifurcations of Vector Fields. — Springer-Verlag 1990.
[17.4] Hale, J.; Koçak, H.: Dynamics and Bifurcations. — Springer-Verlag 1991.
[17.5] Katok, A.; Hasselblatt, B.: Introduction to the Modern Theory of Dynamical Systems. — Cambridge University Press 1995.
[17.6] Kuznetsov, Yu.A.: Elements of Applied Bifurcation Theory. No. 112 in: Applied Mathematical Sciences. — Springer-Verlag 1995.
[17.7] Leonov, G.A.; Reitmann, V.; Smirnova, V.B.: Non-Local Methods for Pendulum-Like Feedback Systems — B. G. Teubner 1987.
[17.8] Mañé, R.: Ergodic Theory and Differentiable Dynamics. — Springer-Verlag 1994.

[17.9] MAREK, M.; SCHREIBER, I.: Chaotic Behaviour of Deterministic Dissipative Systems. — Cambridge University Press 1991.

[17.10] MEDVED', M.: Fundamentals of Dynamical Systems and Bifurcations Theory. — Adam Hilger 1992.

[17.11] PERKO, L.: Differential Equations and Dynamical Systems. — Springer-Verlag 1991.

[17.12] PESIN, YA., B.: Dimension Theory in Dynamical Systems: Contemporary Views and Applications. Chicago Lectures in Mathematics. — The University of Chicago Press 1997.

[17.13] TAKENS, F.: Detecting strange attractors in turbulence. In: Dynamical Systems and Turbulence. Editors: RAND, D. A.; YOUNG, L. S. Lecture Notes in Mathematics 898. — Springer-Verlag 1981, 366–381.

18. Programming, Optimization

[18.1] BÄCK, T.: Evolutionary Algorithms in Theory and Practice. — Oxford University Press 1996.

[18.2] BELLMAN, R.: Dynamic Programming. — Princeton University Press 1957.

[18.3] BERTSEKAS, D.P.: Nonlinear Programming. — Athena Scientific 1999.

[18.4] BEYER, H.-G.: The Theory of Evolution Strategies. Springer-Verlag 2001.

[18.5] CHVATAL, V: Linear Programming. — W.H. Freeman 1983.

[18.6] DANTZIG, G.B.: Linear Programming and Extensions. — Princeton University Press 1998.

[18.7] DENNIS, J.E.; SCHNABEL, R.B.: Numerical Methods for Unconstrained Optimization and Nonlinear Equations. — SIAM 1996.

[18.8] HARZHEIM, L.: Strukturoptimierung – Grundlagen und Anwendungen. — Verlag Harri Deutsch, 2008.

[18.9] KUHN, H.W.: The Hungarian Method for the Assignment Problem. — Naval. Res. Logist. Quart., **2** (1995).

[18.10] MURTY, K.G.: Operations Research: Deterministic Optimization Models. — Prentice Hall 1995.

[18.11] ROCKAFELLAR, R.T.: Convex Analysis. — Princeton University Press 1996.

[18.12] SHERALI, H.D.; BAZARAA, M.S.; SHETTY, C.M.: Nonlinear Programming: Theory and Algorithms. — John Wiley 1993.

[18.13] THAPA, M.N.; DANTZIG, G.B.: Linear Programming 1: Introduction. — Springer-Verlag 1997.

[18.14] WOLSEY, L.A.: Integer Programming. — John Wiley 1998.

19. Numerical Analysis

[19.1] BRENNER, S.C.; SCOTT, L.R.: The Mathematical Theory of Finite Element Methods. — Springer-Verlag 1994.

[19.2] CHAPRA, S.C.; CANALE, R.P.: Numerical Methods for Engineers. — McGraw Hill 1989.

[19.3] COLLATZ, L.: Numerical Treatment of Differential Equations. — Springer-Verlag 1966.

[19.4] DAVIS, P.J.; RABINOWITZ, P: Methods of Numerical Integration. — Academic Press 1984.

[19.5] DE BOOR, C.: A Practical Guide to Splines. — Springer-Verlag 1978.

[19.6] GOLUB, G.; ORTEGA, J.M.: Scientific Computing. — B. G. Teubner 1996.

[19.7] GROSSMANN, CH.; ROOS, H.-G.: Numerik partieller Differentialgleichungen. — B. G. Teubner 1992.

[19.8] HACKBUSCH, W.: Elliptic Differential Equations. — Springer-Verlag 1992.

[19.9] HÄMMERLIN, G.; HOFFMANN, K.–H.: Numerische Mathematik. — Springer-Verlag 1994.

[19.10] HAIRER, E.; NORSETT, S.P.; WANNER, G.: Solving Ordinary Differential Equations. Vol. 1: Nonstiff Problems. Vol. 2: Stiff and Differential Problems. Vol. 3: Algebraic Problems. — Springer-Verlag 1994.

[19.11] HEITZINGER, W.; TROCH, I.; VALENTIN, G.: Praxis nichtlinearer Gleichungen. — C. Hanser Verlag 1984.

[19.12] Kiełbasiński, A.; Schwetlick, H.: Numerische lineare Algebra. Eine computerorientierte Einführung. — Verlag H. Deutsch 1988.
[19.13] Knothe, K.; Wessels, H.: Finite Elemente. Eine Einführung für Ingenieure. — Springer-Verlag 1992.
[19.14] Kress, R.: Numerical Analysis. — Springer-Verlag 1998.
[19.15] Lancaster, P.; Salkauska, S.K.: Curve and Surface Fitting. — Academic Press 1986.
[19.16] Maess, G.: Vorlesungen über numerische Mathematik, Bd. 1, 2. — Akademie-Verlag 1984–1988.
[19.17] Meinardus, G.: Approximation von Funktionen und ihre numerische Behandlung. — Springer-Verlag 1964.
[19.18] Nürnberger, G.: Approximation by Spline Functions. — Springer-Verlag 1989.
[19.19] Pao, Y.-C.: Engineering Analysis. — Springer-Verlag 1998.
[19.20] Quarteroni, A.; Valli, A.: Numerical Approximation of Partial Differential Equations. — Springer-Verlag 1994.
[19.21] Reinsch, Chr.: Smoothing by Spline Functions. — Numer. Math. 1967.
[19.22] Schwarz, H.R.: Methode der finiten Elemente. — B. G. Teubner 1984.
[19.23] Schwarz, H.R.: Numerische Mathematik. — B. G. Teubner 1986.
[19.24] Schwetlick, H.; Kretzschmar, H.: Numerische Verfahren für Naturwissenschaftler und Ingenieure. — Fachbuchverlag 1991.
[19.25] Stoer, J.; Bulirsch, R.: Introduction to Numerical Analysis. — Springer-Verlag 1993.
[19.26] Stroud, A.H.: Approximate Calculation of Multiple Integrals. — Prentice Hall 1971.
[19.27] Törnig, W.: Numerische Mathematik für Ingenieure und Physiker, Bd. 1, 2. — Springer-Verlag 1990.
[19.28] Überhuber, C.: Numerical Computation 1, 2. — Springer-Verlag 1997.
[19.29] Weller, F.: Numerische Mathematik für Ingenieure und Naturwissenschaftler. — Verlag Vieweg 1995.
[19.30] Willers, F.A.: Methoden der praktischen Analysis. — Akademie-Verlag 1951.
[19.31] Zurmühl, R.: Praktische Mathematik für Ingenieure und Physiker. — Springer-Verlag 1984.

20. Computer Algebra Systems

[20.1] Benker, M.: Mathematik mit Mathcad. — Springer-Verlag 1996.
[20.2] Burkhardt, W.: Erste Schritte mit Mathematica. — Springer-Verlag, 2nd ed. 1996.
[20.3] Burkhardt, W.: Erste Schritte mit Maple. — Springer-Verlag, 2nd ed. 1996.
[20.4] Char; Geddes; Gonnet; Leong; Monagan; Watt: Maple V Library, Reference Manual. — Springer-Verlag 1991.
[20.5] Davenport, J.H.; Siret, Y.; Tournier, E.: Computer Algebra. — Academic Press 1993.
[20.6] Gloggengiesser, H.: Maple V. — Verlag Markt & Technik 1993.
[20.7] Gräbe, H.-G.; Kofler, M.: Mathematica. Einführung, Anwendung,Referenz. – Addison-Wesley 1999.
[20.8] Jenks, R.D.; Sutor, R.S.: Axiom. — Springer-Verlag 1992.
[20.9] Kofler, M.: Maple V, Release 4. —Addison Wesley (Deutschland) GmbH 1996.
[20.10] Maeder, R.: Programmierung in Mathematica. — Addison Wesley, 2nd ed. 1991.
[20.11] Trott, M.: The Mathematica Guide Book for Programming. — Springer-Verlag 2004.
[20.12] Trott, M.: The Mathematica Guide Book for Graphics. — Springer-Verlag 2004.
[20.13] Trott, M.: The Mathematica Guide Book for Numerics. — Springer-Verlag 2006.
[20.14] Trott, M.: The Mathematica Guide Book for Symbolics. — Springer-Verlag 2006.
[20.15] Wolfram, S.: The Mathematica Book. — Cambridge University Press 1999.
[20.16] Wolfram, S.: The Mathematica Book. — Wolfram Media 2004.

21. Tables

[21.1] Abramowitz, M.; Stegun, I.A.: Pocketbook of Mathematical Functions. — Verlag H. Deutsch 1984.

[21.2] Apelblat, A.: Tables of Integrals and Series. — Verlag H. Deutsch 1996.

[21.3] Brytschkow, Ju.A.; Maritschew, O.I.; Prudnikow, A.P.: Tabellen unbestimmter Integrale. — Verlag H. Deutsch 1992.

[21.4] Die gesetzlichen Einheiten in Deutschland. PTB-Broschüre. — Phys. Techn. Bundesanstalt.

[21.5] Emde, F.: Tafeln elementarer Funktionen. — B. G. Teubner 1959.

[21.6] Gradstein,I.S.; Ryshik, I.M.: Summen-, Produkt- und Integraltafeln, Bd. 1, 2. — Verlag H. Deutsch 1981.

[21.7] Gröbner, W.; Hofreiter, N.: Integraltafel, Teil 1: Unbestimmte Integrale, Teil 2: Bestimmte Integrale. — Springer-Verlag, Teil 1, 1975; Teil 2, 1973.

[21.8] a) ISO 1000: 11.92-SI units and recommendations for the use of their multiples and of certain other units. b) ISO 31-0 bis ISO 31-XIII.

[21.9] Jahnke, E.; Emde, F.; Lösch, F.: Tafeln höherer Funktionen. — B. G. Teubner 1960.

[21.10] Madelung, E.: Die mathematischen Hilfsmittel des Physikers. — Springer-Verlag 1964.

[21.11] Magnus, W.; Oberhettinger, F.: Formulas and Theorems for the Special Functions of mathematical Physics, 3rd ed. — Springer-Verlag 1966.

[21.12] Müller, H.P.; Neumann, P.; Storm, R.: Tafeln der mathematischen Statistik. — C. Hanser Verlag 1979.

[21.13] Mohr, P.J., Taylor, N.: physics.nist.gov./constants; CODATA Recommended Values of the Fundamental Physical Constants: J. Phys. a. Chem. Ref. Data **28**[6] (1999), Rev. Mod. Phys. **72**[2] (2000).

[21.14] Poljanin, A.D.; Saizew, V.F.: Sammlung gewöhnlicher Differentialgleichungen. — Verlag H. Deutsch 1996.

[21.15] Richtlinie 80/181/EWG des Rates über Einheiten im Meßwesen vom 20.12.1979. (Abl.Nr.**L39/40** vom 15.12.1980, geändert durch Richtlinie 89/617/EWG.)

[21.16] Schüler, M.: Acht– und neunstellige Tabellen zu den elliptischen Funktionen, dargestellt mittels des Jacobischen Parameters q. — Springer-Verlag 1955.

[21.17] Schüler, M.; Gebelein, H.: Acht– und neunstellige Tabellen zu den elliptischen Funktionen. — Springer-Verlag 1955.

[21.18] Schütte, K.: Index mathematischer Tafelwerke und Tabellen. — München 1966.

[21.19] The NIST reference on constants, units and uncertainty http:/www.physics.nist.gov/cuu/Units/current.html

[21.20] The NIST Reference on Constants, Units, and Uncertainty. Fundamental Physical Constants. http://physics.nist.gov./cuu/Constants/archive2002.html Last update see http://physics.nist.gov/cuu/Constants/

[21.21] The NIST Reference on Constants, Units, and Uncertainty. Fundamental Physical Constants. http://physics.nist.gov/constants. Sourc:2010 CODATA recommended values.

22. Handbooks, Guide Books and Reference Books

[22.1] Abramowitz, M.; Stegun, I. A.: Pocketbook of Mathematical Functions. — Verlag H. Deutsch 1984.

[22.2] Baule, B.: Die Mathematik des Naturforschers und Ingenieurs, Bd. 1, 2. — Verlag H. Deutsch 1979.

[22.3] Berendt, G.; Weimar, E.: Mathematik für Physiker, Bd. 1, 2. — VCH 1990.

[22.4] Bourbaki, N.: The Elements of Mathematics, Vols. 1ff. — Springer-Verlag 1990ff.

[22.5] Bronstein, J.N.; Semendjajew, K.A.: Taschenbuch der Mathematik. — B. G. Teubner 1989, 24., Auflage; Verlag H. Deutsch 1989.

[22.6] Bronstein, J.N.; Semendjajew, K.A.: Taschenbuch der Mathematik, Ergänzende Kapitel. — Verlag H. Deutsch 1991.

[22.7] Bronstein, J.N.; Semendjajew, K.A.; Musiol, G.; Muehlig, H.: Handbook of Mathematics. — Springer-Verlag 2004, 4th Ed.

[22.8] Bronstein, J.N.; Semendjajew, K.A.; Musiol, G.; Muehlig, H.: Taschenbuch der Mathematik. — Verlag Harri Deutsch 2012, 9th german Ed.
[22.9] Dreszer, J.: Mathematik. – Handbuch für Technik und Naturwissenschaft. — Verlag H. Deutsch 1975.
[22.10] Gellert, W.; Kästner, H.; Neuber, S. (Eds.): Fachlexikon ABC Mathematik. — Verlag H. Deutsch 1978.
[22.11] Fichtenholz, G.M.: Differential- und Integralrechnung, Bd. 1, 3. — Verlag H. Deutsch 1994.
[22.12] Gellert, W.; Küstner, H.; Hellwich, M.; Kästner, H. (Eds.): Kleine Enzyklopädie Mathematik. — Verlag Enzyklopädie, Leipzig 1965.
[22.13] Joos, G.; Richter, E.W.: Höhere Mathematik für den Praktiker. — Verlag H. Deutsch 1994.
[22.14] Mangoldt, H. v.; Knopp, K.; Hrg. F. Lösch: Einführung in die höhere Mathematik, Bd. 1–4. — S. Hirzel-Verlag 1989.
[22.15] Margenau, H.; Murphy, G.M.: The Mathematics of Physics and Chemistry, Vols. 1, 2. — Van Nostrand 1956; Verlag H. Deutsch 1965, 1967.
[22.16] Papula, L.: Mathematik für Ingenieure, Bd. 1–3. — Verlag Vieweg 1994–1996.
[22.17] Plaschko, P.; Brod, K.: Höhere mathematische Methoden für Ingenieure und Physiker. — Springer-Verlag 1989.
[22.18] Precht, M.; Voit, K.; Kraft, R.: Mathematik für Nichtmathematiker, Bd. 1, 2. — Oldenbourg-Verlag 1991.
[22.19] Rothe, R.: Höhere Mathematik für Mathematiker, Physiker, Ingenieure, Teil I–IV. — B. G. Teubner 1958–1964.
[22.22] Schmutzer, E.: Grundlagen der theoretischen Physik, Bd. 1, 4. — Deutscher Verlag der Wissenschaften 1991.
[22.23] Smirnow, W.I.: Lehrbuch der höheren Mathematik, Bd. 1–5. — Verlag H. Deutsch 1994.
[22.24] Zeidler, E. (Ed.): Teubner Taschenbuch der Mathematik. — B. G. Teubner, Teil 1, 1996; Teil 2, 1995.

23. Encyclopedias

[23.1] Eisenreich, G., Sube, R.: Wörterbuch Mathematik, Vols. 1, 2 (English, German, French, Russian) — Verlag Technik 1982; Verlag H. Deutsch 1982.
[23.2] Encyclopaedia Britannica.
[23.3] Encyclopaedia of Mathematical Sciences (Transl. from the Russian). — Springer-Verlag 1990.
[23.4] Encyclopaedia of Mathematics (Revised Transl. from the Russian). — Kluwer 1987–1993.

Index

I.N. Bronshtein et al., *Handbook of Mathematics*,
DOI 10.1007/978-3-662-46221-8

MATHEMATICAL SYMBOLS

Relational Symbols

$=$	equal to	$\approx$	approximately equal to	$\leq$	less than or equal to
$\equiv$	identically equal to	$<$	less than	$\geq$	greater than or equal to
$:=$	equal to by definition	$>$	greater than	$\neq$	unequal to, different from
$\ll$	much less than	$\gg$	much greater than	$\hat{=}$	corresponding to
$\prec$	partial order relation	$\succ$	partial order relation		

Greek Alphabet

$A\ \alpha$	Alpha	$B\ \beta$	Beta	$\Gamma\ \gamma$	Gamma	$\Delta\ \delta$	Delta	$E\ \varepsilon$	Epsilon	$Z\ \zeta$	Zeta
$H\ \eta$	Eta	$\Theta\ \theta\ \vartheta$	Theta	$I\ \iota$	Iota	$K\ \kappa$	Kappa	$\Lambda\ \lambda$	Lambda	$M\ \mu$	Mu
$N\ \nu$	Nu	$\Xi\ \xi$	Xi	$O\ o$	Omicron	$\Pi\ \pi$	Pi	$P\ \rho$	Rho	$\Sigma\ \sigma$	Sigma
$T\ \tau$	Tau	$\Upsilon\ \upsilon$	Upsilon	$\Phi\ \varphi$	Phi	$X\ \chi$	Chi	$\Psi\ \psi$	Psi	$\Omega\ \omega$	Omega

Constants

const	constant amount (constant)	$C = 0.57722\ldots$	Euler constant
$\pi = 3.14159\ldots$	ratio of the perimeter of the circle to the diameter	$e = 2.71828\ldots$	base of the natural logarithms

Algebra

A, B	propositions	$\neg A, \overline{A}$	negation of the proposition A
$A \wedge B, \sqcap$	conjunction, logical AND	$A \vee B, \sqcup$	disjunction, logical OR
$A \Rightarrow B$	implication, IF A THEN B	$A \Leftrightarrow B$	equivalence, A IF AND ONLY IF B
$A, B, C, \ldots$	sets	$\mathbb{N}$	set of natural numbers
$\overline{A}$	closure of the set A or complement of A with respect to a universal set	$\mathbb{Z}$ $\mathbb{Q}$	set of the integers set of the rational numbers
$A \subset B$	A is a proper subset of B	$\mathbb{R}$	set of the real numbers
$A \subseteq B$	A is a subset of B	$\mathbb{R}_+$	set of the positive real numbers
$A \setminus B$	difference of two sets	$\mathbb{R}^n$	n-dimensional Euclidean vector space
$A \triangle B$	symmetric difference	$\mathbb{C}$	set of the complex numbers
$A \times B$	Cartesian product	$R \circ S$	relation product
$x \in A$	x is an element of A	$x \notin A$	x is not an element of A
card A	cardinal number of the set A	$\emptyset$	empty set, zero set
$A \cap B$	intersection of two sets	$\bigcap_{i=1}^{n} A_i$	intersection of n sets A_i
$A \cup B$	union of two sets	$\bigcup_{i=1}^{n} A_i$	union of n sets A_i
$\forall\, x$	for all elements x	$\exists\, x$	there exists an element x
$\{x \in X : p(x)\}$	subset of all x from X of the property $p(x)$	$\{x : p(x)\}$, $\{x \mid p(x)\}$	set of all x with the property $p(x)$
$T\colon X \longrightarrow Y$	mapping T from the space X into the space Y	$\cong$ $\sim_R$	isomorphy of groups equivalence relation
$\oplus$	residue class addition	$\odot$	residue class multiplication
$H = H_1 \oplus H_2$	orthogonal decomposition of space H	$\mathbf{A} \otimes \mathbf{B}$	Kronecker product
supp	support	iff	if and only if

sup M supremum: least upper bound of the non-empty set M ($M \subset \mathbb{R}$) bounded above

inf M infimum: greatest lower bound of the non-empty set M ($M \subset \mathbb{R}$) bounded below

I.N. Bronshtein et al., *Handbook of Mathematics*,
DOI 10.1007/978-3-662-46221-8

$[a, b]$	closed interval, i.e.,	$\{x \in \mathbb{R} \colon a \le x \le b\}$
$(a, b),]a, b[$	open interval, i.e.,	$\{x \in \mathbb{R} \colon a < x < b\}$
$(a, b],]a, b]$	interval open from left, i.e,	$\{x \in \mathbb{R} \colon a < x \le b\}$
$[a, b), [a, b[$	interval open from right, i.e.,	$\{x \in \mathbb{R} \colon a \le x < b\}$

$\operatorname{sign} a$	sign of the number a, e.g., $\operatorname{sign}(\pm 3) = \pm 1$, $\operatorname{sign} 0 = 0$
$\mid a \mid$	absolute value of the number a
$\lfloor a \rfloor$	integer, greater or equal to a (compare $[x]$, p. C)
a^m	a to the power m, a to the m-th
$\sqrt{a}$	square root of a
$\sqrt[n]{a}$	n-th root of a
$\log_b a$	logarithm of the number a to the base b, e.g., $\log_2 32 = 5$
$\log a$	decimal logarithm (base 10) of the number a, e.g., $\lg 100 = 2$
$\ln a$	natural logarithm (base e) of the number a, e.g., $\ln e = 1$

$a \mid b$	a is a divisor of b, a devides b, the ratio of a to b
$a \nmid b$	a is not a divisor of b
$a \equiv b \bmod m$, $a \equiv b(m)$	a is congruent to b modulo m, i.e., $b - a$ is divisible by m
g.c.d.$(a_1, a_2, \ldots, a_n)$	greatest common divisor of $a_1, a_2, \ldots, a_n$
l.c.m.$(a_1, a_2, \ldots, a_n)$	least common multiple of $a_1, a_2, \ldots, a_n$
$\binom{n}{k}$	binomial coefficient, n choose k
$\left(\frac{a}{b}\right)$	Legendre symbol
$n! = 1 \cdot 2 \cdot 3 \cdot \ldots \cdot n$	factorial, e.g., $6! = 1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \cdot 6 = 720$; specially: $0! = 1! = 1$
$(2n)!! = 2 \cdot 4 \cdot 6 \cdot \ldots \cdot (2n) = 2^n \cdot n!$;	in particular: $0!! = 1!! = 1$
$(2n+1)!! = 1 \cdot 3 \cdot 5 \cdot \ldots \cdot (2n+1)$	

$\mathbf{A} = (a_{ij})$	matrix A with elements a_{ij}	$\mathbf{A}^{\mathrm{T}}$	transposed matrix
$\mathbf{A}^{-1}$	inverse matrix	$\mathbf{A}^{\mathrm{H}}$	adjoint matrix
$\mathbf{E} = (\delta_{ij})$	unit matrix	$\mathbf{0}$	zero matrix
rank	rank of a matrix	trace	trace of a matrix

$\det \mathbf{A}$, D	determinant of the square matrix A
δ_{ij}	Kronecker symbol: $\delta_{ij} = 0$ for $i \ne j$ and $\delta_{ij} = 1$ for $i = j$
$\underline{\mathbf{a}}$	column vector in $\mathbb{R}^n$
$\underline{\mathbf{a}}^{\mathbf{0}}$	unit vector in the direction of (parallel to) $\underline{\mathbf{a}}$
$\Vert \underline{\mathbf{a}} \Vert$	norm of $\underline{\mathbf{a}}$
$\vec{\mathbf{a}}, \vec{\mathbf{b}}, \vec{\mathbf{c}}$	vectors in $\mathbb{R}^3$
$\vec{\mathbf{i}}, \vec{\mathbf{j}}, \vec{\mathbf{k}}$ $\quad \vec{\mathbf{e}}_x, \vec{\mathbf{e}}_y, \vec{\mathbf{e}}_z$	basis vectors (orthonormed) of the Cartesian coordinate system
a_x, a_y, a_z	coordinates (components) of the vector $\vec{\mathbf{a}}$
$\mid \vec{\mathbf{a}} \mid$	absolute value, length of the vector $\vec{\mathbf{a}}$
$\alpha \underline{\mathbf{a}}$	multiplication of a vector by a scalar
$\vec{\mathbf{a}} \cdot \vec{\mathbf{b}}, \vec{\mathbf{a}}\vec{\mathbf{b}}, (\vec{\mathbf{a}}\vec{\mathbf{b}})$	scalar product, dot product
$\vec{\mathbf{a}} \times \vec{\mathbf{b}}, [\vec{\mathbf{a}}\vec{\mathbf{b}}]$	vector product, cross product
$\vec{\mathbf{a}}\vec{\mathbf{b}}\vec{\mathbf{c}} = \vec{\mathbf{a}} \cdot (\vec{\mathbf{b}} \times \vec{\mathbf{c}})$	parallelepipedal product, mixed product (triple scalar product)
$\underline{\mathbf{0}}, \vec{\mathbf{0}}$	zero vector
$\boldsymbol{T}$	tensor
$G = (V, E)$	graph with the set of vertices V and the set of edges E

Geometry

$\perp$	orthogonal (perpendicular)	$\|$	parallel
$\#$	equal and parallel	$\sim$	similar, e.g., $\triangle ABC \sim \triangle DEF$; proportional
$\triangle$	triangle	$\measuredangle$	angle, e.g., $\measuredangle ABC$
$\frown$	arc segment, e.g., $\widehat{AB}$ the arc between A and B	rad	radian

$^\circ$ degree, $'$ minute, $''$ second — as measure of angle and circular arc, e.g., $32^\circ\ 14'\ 11.5''$

gon — measure in grades ($360^\circ = 400\,\text{gon}$, see p. 131 and table3.5, p. 146)

$\overline{AB}$ — the line segment between A and B

$\overrightarrow{AB}$ — the directed line segment from A to B, the vector from A to B

Complex Numbers

i (sometimes j)	imaginary unit ($\mathrm{i}^2 = -1$)	I	imaginary unit in computer algebra
$\mathrm{Re}\,(z)$	real part of the number z	$\mathrm{Im}\,(z)$	imaginary part of the number z
$\lvert z\rvert$	absolute value of z	$\arg z$	argument of the number z
$\bar{z}$ or z^*	complex conjugate of z, e.g., $z = 2+3\mathrm{i}$, $\bar{z} = 2-3\mathrm{i}$	$\mathrm{Ln}\, z$	logarithm (natural) of a complex number z

Trigonometric Functions, Hyperbolic Functions

sin	sine	cos	cosine
tan	tangent	cot	cotangent
sec	secant	cosec	cosecant
arcsin	principal value of arc sine (sine inverse)	arccos	principal value of arc cosine (cosine inverse)
arctan	principal value of arc tangent (tangent inverse)	arccot	principal value of arc cotangent (cotangent inverse)
arcsec	principal value of arc secant (secant inverse)	arccosec	principal value of arc cosecant (cosecant inverse)
sinh	hyperbolic sine	cosh	hyperbolic cosine
tanh	hyperbolic tangent	coth	hyperbolic cotangent
sech	hyperbolic secant	cosech	hyperbolic cosecant
Arsinh	area–hyperbolic sine	Arcosh	area–hyperbolic cosine
Artanh	area–hyperbolic tangent	Arcoth	area–hyperbolic cotangent
Arsech	area–hyperbolic secant	Arcosech	area–hyperbolic cosecant

Analysis

$\lim\limits_{n\to\infty} x_n = A$	A is the limit of the sequence (x_n). We also write $x_n \to A$ as $n \to \infty$; e.g., $\lim\limits_{n\to\infty} \left(1+\frac{1}{n}\right)^n = e$
$\lim\limits_{x\to a} f(x) = B$	B is the limit of the function $f(x)$ as x tends to a
$[x] = \mathrm{int}(x)$ $(\mathrm{entier}(x))$	greatest integer less or equal to x (compare p. B)
$f = o(g)$ for $x \to a$	Landau symbol "small o" means: $f(x)/g(x) \to 0$ as $x \to a$
$f = O(g)$ for $x \to a$	Landau symbol "big O" means: $f(x)/g(x) \to C$ ($C = \text{const}$, $C \neq 0$) as $x \to a$
$\sum\limits_{i=1}^{n}$, $\sum_{i=1}^{n}$	sum of n terms for i equals 1 to n
$\prod\limits_{i=1}^{n}$, $\prod_{i=1}^{n}$	product of n factors for i equals 1 to n
$f(\ )$, $\varphi(\ )$	notation for a function, e.g., $y = f(x)$, $u = \varphi(x, y, z)$
Δ	difference or increment, e.g., $\Delta\, x$ (delta x)
d	differential, e.g., dx (differential of x)
$\dfrac{d}{dx}, \dfrac{d^2}{dx^2}, \ldots, \dfrac{d^n}{dx^n}$	determination of the first, second, ..., n-th derivative with respect to x

$\left.\begin{array}{l} f'(x), f''(x), f'''(x), \\ f^{(4)}(x), \ldots, f^{(n)}(x) \\ \text{or} \\ \dot{y}, \ddot{y}, \ldots, y^{(n)} \end{array}\right\}$	first, second,..., n-th derivative of the function $f(x)$ or of the function y
$\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial^2}{\partial x^2}, \ldots$	determination of the first, second, ..., n-th partial derivative
$\frac{\partial^2}{\partial x \partial y}$	determination of the second partial derivative first with respect to x, then with respect to y
$f_x, f_y, f_{xx}, f_{xy}, f_{yy}, \ldots$	first, second, ... partial derivative of function $f(x, y)$
D	differential operator, e.g., $Dy = y', D^2y = y''$
grad	gradient of a scalar field ($\operatorname{grad} \varphi = \nabla \varphi$)
div	divergence of a vector field ($\operatorname{div} \vec{\mathbf{v}} = \nabla \cdot \vec{\mathbf{v}}$)
rot	rotation or curl of a vector field ($\operatorname{rot} \vec{\mathbf{v}} = \nabla \times \vec{\mathbf{v}}$)
$\nabla = \frac{\partial}{\partial x}\vec{\mathbf{i}} + \frac{\partial}{\partial y}\vec{\mathbf{j}} + \frac{\partial}{\partial z}\vec{\mathbf{k}}$	nabla operator, here in Cartesian coordinates (also called the Hamiltonian differential operator, not to be confused with the Hamilton operator in quantum mechanics)
$\Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$	Laplace operator
$\frac{\partial \varphi}{\partial \vec{\mathbf{a}}}$	directional derivative, i.e., derivative of a scalar field φ into the direction $\vec{\mathbf{a}}$: $\frac{\partial \varphi}{\partial \vec{\mathbf{a}}} = \vec{\mathbf{a}} \cdot \operatorname{grad} \varphi$
$\int\limits_a^b f(x)dx,$	definite integral of the function f between the limits a and b
$\int\limits_{(C)} f(x, y, z)ds$	line integral of the first kind with respect to the space curve C with arclength s
$\oint\limits_{(C)} f(x, y, z)\, ds$	integral along a closed curve (circulatory integral)
$\iint\limits_{(S)} f(x, y)dS = \iint\limits_{(S)} f(x, y)\, dx\, dy$	double integral over a planar region S
$\int\limits_{(S)} f(x, y, z)\, dS = \iint\limits_{(S)} f(x, y, z)dS$	surface integral of the first kind over a spatial surface S (see (8.151b), p. 534)
$\int\limits_{(V)} f(x, y, z)\, dV = \iiint\limits_{(V)} f(x, y, z)\, dx\, dy\, dz$	triple integral or volume integral over the volume V
$\left.\begin{array}{l} \oint\limits_{(S)} U(\vec{\mathbf{r}})\, \mathbf{d}\vec{\mathbf{S}} = \oint\!\!\!\oint\limits_{(S)} U(\vec{\mathbf{r}})\, \mathbf{d}\vec{\mathbf{S}} \\ \oint\limits_{(S)} \vec{\mathbf{V}}(\vec{\mathbf{r}}) \cdot \mathbf{d}\vec{\mathbf{S}} = \oint\!\!\!\oint\limits_{(S)} \vec{\mathbf{V}}(\vec{\mathbf{r}}) \cdot \mathbf{d}\vec{\mathbf{S}} \\ \oint\limits_{(S)} \vec{\mathbf{V}}(\vec{\mathbf{r}}) \times \mathbf{d}\vec{\mathbf{S}} = \oint\!\!\!\oint\limits_{(S)} \vec{\mathbf{V}}(\vec{\mathbf{r}}) \times \mathbf{d}\vec{\mathbf{S}} \end{array}\right\}$	surface integrals over a closed surface in vector analysis
$A = \max!$	expression A is to be maximized, similarly min!, extreme!
$A = \max$	expression A is maximal, similarly min, extreme.